Collaborative Curation in Social Production Communities

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Shyong (Tony) K. Lam

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

John T. Riedl, Advisor

August, 2012

Acknowledgements

This dissertation would not have been possible without the support, guidance, and contributions of the numerous friends, colleagues, and collaborators who I have met along the way.

First and foremost, I thank my advisor John Riedl for sharing his sage-like wisdom, his insights in being an effective scientist, and his patience throughout the years. Even as he was enduring his own trials and tribulations, he always managed to find the time and energy to work with me and to challenge me to better myself, and for that, I am forever grateful.

A special thanks to Dr. Loren Terveen, Dr. Yongdae Kim, and Dr. Gedas Adomavicius for serving on my thesis committee and providing valuable feedback about my work throughout the process.

I thank Steve Lawrence, David Pennock, Ed Chi, and Elizabeth Churchill for providing me with internship opportunities that allowed me to enrich my skills and broaden my views of research in industry.

I would also like to acknowledge the National Science Foundation for the financial support that made my work possible.

GroupLens Research is filled with talented and brilliant people, and being surrounded by all of them made it a wonderful environment to do my work. To Sean, Nathan, DanCo, Shilad, Reid, Aaron, Katie, Morten, Anu, Joe, and the dozens of others: thank you. I am especially grateful to Jon Herlocker, who saw something in me when I was a freshman and brought me into GroupLens. Jon was my introduction to the world of academic research, and without his influence and mentorship, this dissertation may never have happened.

Finally, I would not have been able to complete this journey without the endless moral support of my friends and my family. <3

Dedication

To my parents.

Abstract

The use of *social production communities* (SPCs) has become a common approach for building information repositories such as Wikipedia, Yahoo! Answers, and YouTube. In these systems, communities of users collaborate to produce a shared repository of information. We define *collaborative curation* as the tasks performed by these communities, and the processes, workflows, and policies that guide how users work together. This thesis seeks to study the implications of different curation processes, and the challenges that SPCs face in constructing information repositories. Our goal is to better understand the growth and evolution of SPC information repositories so that we can inform the design of SPCs.

The first part of this thesis focuses on collaborative curation practices at a high level to learn about the design space of curation mechanisms and the impact that different mechanisms have on the evolution of SPCs. We begin with an analysis of Wikipedia's curation practices, studying how Wikipedia's editors decide which articles merit inclusion in the encyclopedia, and how the encyclopedia has grown over the years. We then conduct a user study using the MovieLens recommender system to compare two typical curation mechanisms – a wiki-like process, and a social voting process – in how they affect the growth of MovieLens' movie database.

In the second part of this thesis, our focus shifts to challenges that SPCs face in collaborative curation. We start by looking at how skews in group composition can influence collaborative curation. SPCs typically rely on the efforts of self-formed and self-organized volunteer groups. Such groups may differ from the larger user community or from the general populace on multiple dimensions, including demographics, attitudes, and experience. We conduct two studies to study these differences in the context of Wikipedia. At the small-scale level, we examine how composition skews in small working groups can affect curation decision quality; at the large-scale level, we explore an apparent gender disparity amongst Wikipedia's community of editors. We close with an analysis of a type of malicious deviant behavior where users submit false data to an SPC in an attempt to manipulate choices made by fellow users.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Throughout recorded history, humankind has been remarkably successful in constructing vast repositories of information. Several millennia ago in Egypt, scholars collected hundreds of thousands of ancient papyrus scrolls for the Library of Alexandria's archives. During the Middle Ages, theologists, philosophers, and rhetoricians produced vast compendiums of knowledge such as the *Etymologiae* and the *Suda*, which are among the world's earliest encyclopedias. Over the past two centries, an army of historians, biologists, anthropologists, and scientists in other disciplines have amassed and maintained millions of artifacts and specimens in the Smithsonian Institution's collections. Even the storied internet giant Yahoo! began life not as a search or media company, but as *Jerry and David's Guide to the World Wide Web*, a hand-picked list of links to interesting web sites and pages that was built by founders Jerry Yang and David Filo, who were, at the time, doctoral candidates at Stanford University.[1]

Traditionally, these repositories have been painstakingly constructed and carefully maintained by professionals who are expected to have substantial domain expertise. Awareness of the maintainers' credentials can instill a sense of quality and trustworthiness in an information repository [27]. This type of expert oversight has been used to build numerous libraries, academic journals, museums, encyclopedias, and newspapers that are invaluable resources for information seekers.

However, this way of building information repositories has a substantial drawback: cost. Even as the proliferation of cheap commodity computing and internet access continues to drive

[1] http://docs.yahoo.com/info/misc/history.html

down costs of digital storage, communication, and dissemination of information, a major expense remains: expert labor. Professionals and experts typically prefer not to work pro bono, and therefore, the ability to develop and maintain these repositories is often limited to well-financed organizations such as governments and successful companies. Even then, some institutions have been struggling to make ends meet; for instance, the American Library Association reports that public library systems in over half of the United States have had their budgets cut substantially between 2008 and 2011, resulting in numerous branch closures [7].

1.1 A New Way Forward

The advent of the internet offered a radical approach to building information repositories. Just as the internet opens up unprecedented levels of access to information, it also allows people to easily share knowledge, opinions, and media. By harnessing the collective effort and intelligence of millions of internet users, some systems have successfully built information repositories without the need to require paid professionals to do all of the work. In this thesis, we refer to such systems as *social production communities*, or SPCs. [2]

At first blush, such an approach might seem absurd. After all, how can professionals be replaced by a group of volunteers who may not have any verifiable qualifications? Could a resource produced via "the wisdom of crowds" [149] be at all trustworthy? Computer visionary Jaron Lanier has been particularly skeptical and critical of this paradigm; in [92], he writes:

> "A fashionable idea in technical circles is that quantity not only turns into quality at some extreme of scale, but also does so according to principles we already understand. Some of my colleagues think a million, or perhaps a billion, fragmentary insults will eventually yield wisdom that surpasses that of any well-thought-out essay, so long as sophisticated secret statistical algorithms recombine the fragments. I disagree. A trope from the early days of computer science comes to mind: garbage in, garbage out."

[2] These sorts of systems have been associated with numerous names in the mass media as well as the research literature, including *Web 2.0*, commons-based peer production, community-maintained artifacts of lasting value, and online communities. Though each of these names may refer to somewhat different classes or different aspects of these systems, such nuances and distinctions are immaterial to the work presented in this thesis. By "social production community," we mean any online community whose users are collectively building and maintaining a shared repository of information.

Despite these reservations, SPC-maintained information repositories have become commonplace on the internet, and are used by millions of users daily. For instance, Wikipedia has become the world's largest encyclopedia, and Q&A site Yahoo! Answers has helped people get over a billion answers to their questions.[3] Even with limited (or zero) oversight by paid professionals, such systems are apparently of sufficiently high-quality to be useful resources, and in some cases, have supplanted professionally-maintained resources. Consider the open-source software ecosystem, which produces a vast repository of software that is maintained largely by unpaid volunteers. Its crown jewel is Linux, a freely-available computer operating system that has complemented or replaced commercial operating systems in many organizations around the world [152].

SPCs rely on their community of users to work together to produce a useful information repository. While this idea is simple and powerful, and has been used to great effect in real-world systems, actually building a successful SPC is difficult. Motivating individuals to participate in constructive ways is a key issue; the *free rider problem* and the *tragedy of the commons* dilemma [61] suggest that while many people stand to benefit from successful SPCs, they may not be quite as willing to contribute. Also, since those who contribute will not necessarily be experts in the relevant subject area(s), quality control can be a major concern.

As a result, many SPCs do ultimately fail [4, 37, 166, 141, 33, 132]. Some systems wither away due to simple lack of community participation, while others are marred by unwanted forms of participation such as spamming, trolling, excessive in-fighting, or low-quality contributions.

1.2 Collaborative Curation

One important factor in determining whether an SPC is successful is how its repository of information is managed. The tasks that SPC participants perform as they manage a repository are, in many ways, similar to those that curators are responsible for. In its code of ethics for curators, the American Association of Museums' Curators Committee lists several tasks that curators are responsible for, including [111]:

- Make recommendations for acquiring and deaccessioning objects in the museum collection.

[3] http://yanswersblog.com/index.php/archives/2010/05/03/1-billion-answers-served/

- Assume responsibility for the overall care and development of the collection [...]
- Advocate for and participate in the formulation of institutional policies and procedures for the care of the collection [...]
- Perform research to identify materials in the collection and to document their history.
- Develop and organize exhibitions.

Similarly, in an SPC, the community of contributors can be seen as working collaboratively to curate a repository of information. They may help decide what items belong in the repository, propose and add new items to it, modify or augment existing items to improve them, or delete items that violate norms or rules. In some SPCs, they may also be able to help write or refine policies related to the ongoing maintenance of the information repository. We define the term *collaborative curation* as the sum of all these information repository management tasks and processes in the context of an SPC.

One primary difference between traditional curation and collaborative curation is the massively collaborative and often decentralized nature of the latter. While traditional curation typically involves a single person or a small close-knit group, collaborative curation is performed by hundreds if not thousands of people who are working simultaneously in numerous areas across the SPC's repository.

To manage the community's efforts, SPC designers need to establish processes that define and shape the ways that people collaborate to curate the repository together. Three key high-level design elements of collaborative curation processes in an SPC are as follows:

Contribution Workflow. All SPCs have a workflow for managing how contributors create, modify, or delete items in the information repository. The workflow determines how people find tasks to do, who is able to propose each type of contribution, and what must occur before a proposed contribution is applied to the repository.

Some SPCs require submissions to undergo a formal review process involving peers or possibly staff members. Such processes are typically intended for quality control and help verify that contributions adhere to established quality standards and content policies. Other SPCs are less stringent and utilize post-hoc peer review as their primary quality control mechanism, relying on users to be vigilant while perusing the repository and to help fix problems stemming from low-quality or counterproductive contributions.

Conflict Resolution. In any SPC, numerous decisions must be made regarding issues such as determining whether contributions are appropriate, interpreting and applying policies, upholding social norms, responding to deviant behavior, and so on. Conflicts and disagreements between individuals or subgroups are inevitable.

SPCs may provide various dispute resolution procedures and venues to guide involved parties in working through conflicts and making decisions amongst themselves, or in escalating and drawing attention to situations that have reached an impasse. A wide variety of conflict resolution strategies exist, including simply providing guidelines and advice about decision-making, establishing venues for community discussions aimed at finding a group consensus, and perhaps in extreme cases, organizing formal proceedings that resemble arbitration hearings or courtroom trials.

Governance. The rules and policies of an SPC define what is expected from the community and how curation works. In addition to specifying what the contribution review and conflict resolution procedures are, the rules and policies also codify things such as:

- Information repository scope: What is the goal of the SPC? What types of items belong in the information repository, and what types do not?
- Community structure: What leadership roles are available to community members? How do people get selected for those roles, and what privileges and responsibilities do people in these roles have?
- Behavioral expectations: What types of user behaviors are acceptable or unacceptable? Is anonymous or pseudonymous participation allowed? How are the rules enforced?

Some SPC operators may opt to retain unilateral control of governance. Others may provide the community with ways to provide input and feedback, or to have some level of control over the rules and policies.

Existing SPCs exhibit a wide range of collaborative curation practices. Some are quite open and have few rules or criteria for inclusion. For example, the photo sharing site Flickr, which has accumulated billions of photos from its community of users, encourages its members to upload any of their photos. Besides honoring legal obligations (i.e., copyright infringement claims), Flickr imposes few guidelines regarding quality or subject matter. New photos are immediately

made available on the site without an approval or vetting process, and are subject to editorial review only if others in the community lodge complaints about the photo.

As an example of a site that takes a very different approach at curation, consider the Internet Movie Database (IMDb), which offers its users the ability to submit movie information such as film titles, actors, and memorable quotes to its vast online repository. However, before they are actually added to the repository, all submissions are subject to an editorial review process in which IMDb's paid staff examine and evaluate each proposed modification according to a comprehensive set of guidelines. The process is time-consuming, and according to the official FAQ, it may take *several weeks* to add a film to IMDb. [4] This is an example of a curation process that is only partially community-driven and is subject to substantial editorial control and oversight.

In between these two examples lie numerous approaches to curation that grant varying levels of control and governance to users, and that offer different workflows for curating information repositories. Table 1.1 describes a wide range of SPCs and the collaborative curation processes that they utilize.

This thesis seeks to study the implications of these different collaborative curation processes, and the challenges that SPCs face in constructing information repositories. Since the design space of SPCs and collaborative curation processes is very large, our explorations will not be exhaustive. We choose to focus on areas that cover interesting parts of the space. We look at SPC designs that are representative of those that have been successful in existing systems, and at collaborative curation challenges that real-world SPCs face. Our goal is to better understand the growth and evolution of SPC information repositories so that we can help SPC designers and maintainers more effectively achieve their goals and build successful communities.

1.3 Contributions

In the first part of our work, we study collaborative curation practices at a high level to learn more about the design space and the broad impact that different design decisions can have on the growth and evolution of SPCs. First, we perform a detailed analysis of the collaborative curation practices in one of the world's largest SPCs, the English Wikipedia. We examine the policies and processes that Wikipedians use to decide which articles merit inclusion in the

[4] http://www.imdb.com/help/show_leaf?wheresmytitle

SPC	Contribution Workflow	Conflict Resolution	Governance
Wikipedia	Users (mostly) have the ability to create and modify any encyclopedia article without need for review before contributions are visible. Article deletions can be proposed by any user, and must have support from the community in order to occur. Informal post-hoc peer review is used for quality control.	Many conflicts are resolved via discussion and consensus-finding activities (e.g., straw polls). Administrators may revoke users' editing privileges in cases of substantial disruption. Community-operated mediation and arbitration venues are available for serious conflicts.	System policies and rules are written and enforced by the community. Administrator, mediator, and arbitrator privileges are granted through community-defined voting or appointment processes. The Wikimedia Foundation, which operates Wikipedia, maintains a hands-off approach to most matters.
Nupedia [a]	All users could participate, but authors and reviewers were expected to have some level of expertise in the appropriate field(s). Submitted articles were subject to multiple rounds of peer review and copyediting before being considered ready for public access.	Disputes that could not be resolved within an editorial group were resolved by the chief copyeditor or by the editor-in-chief.	Nupedia's editor-in-chief had control over policy as well as content. However, category-specific editorial groups were expected to determine content formats and guidelines for articles under their purview.
StackOverflow (Q&A for programmers)	All users may ask and answer questions. Submitted questions and answers published to the site without formal review. Informal post-hoc peer review is used for quality control. By making quality sumissions, users may earn additional curation privileges such as editing others' questions and answers and participating in question deletion processes.	Users with sufficient reputation may cast negative votes to reduce the visibility of poor questions or flawed answers. Users may also earn the privilege to flag inappropriate questions or answers to request moderator or staff review. A "meta" area is available for discussing broader issues.	The various curation mechanisms are defined and controlled by site operators. Volunteer moderators are appointed by site staff or are elected via popular vote, and are responsible for carrying out privileged SPC management tasks such as deleting inappropriate content, merging related questions, and suspending users who are being disruptive.
Reddit, Digg (social news)	Users may submit web links and descriptions to the repository, but all submissions are subject to a social voting process that determines which links will appear on the front page. Users may view submitted links and vote for ones that they believe are interesting enough to appear on the front page.	The voting process allows negative votes for low-quality submissions and comments. Inappropriate submissions can be reported to the system for editorial review.	Site operators control the vote tabulation algorithms and overarching content policies that determine what submissions are inappropriate.
Apache Web Server (open source) [b]	Voting processes existed for different types of proposals or contributions. For example, source code changes needed three supporting votes and no vetos.	Conflicts are resolved via informal email discussions and the voting processes.	Only members of the Apache Group could cast binding votes. Becoming a member required unaminous support from existing members.
Internet Movie Database	Users may submit new or updated movie information. All submissions are reviewed by site staff before proposed changes are made visible on the site.	Site staff is responsible for making all decisions.	Site operators have unilateral control over policy.
Flickr, YouTube, Facebook, Twitter	Users have full control in curating their own collections of information, but no ability to modify others' collections. No review process is required before new contributions are made visible. [c]	Users can report instances of inappropriate contributions (e.g., spam or sexually explicit material) to the system, and can use blocking functionality to avoid undesired communication.	Site operators have unilateral control over policy.

Table 1.1: Examples of collaborative curation processes in real-world SPCs.

[a] Nupedia is a defunct predecessor to Wikipedia that, like Wikipedia, was founded and operated by Jimmy Wales and Larry Sanger [1].

[b] These procedures are described in [39] and characterize the Apache development group's processes as of the late 1990s.

[c] Some of these SPCs employ staff to spot-check contributions, e.g. [148].

encyclopedia, and study how these processes have affected the growth of Wikipedia over the years. Then, moving from offline data analysis to online experimentation, we explore how different mechanism design decisions affect user behavior and repository growth in an existing SPC, MovieLens.

We then turn to an exploration of several challenges that occur during collaborative curation in real-world SPCs. We begin by studying how differences in group composition can influence collaborative curation. Because the volunteer groups that perform various curation tasks are often self-organized, they are not necessarily representative of the larger community of users or of the general populace. In turn, the tasks that they perform and the decisions that they make in curating the SPC may be affected, possibly leading to abnormalities in the repository's growth and evolution. We look at both small working groups and entire populations on Wikipedia, identifying imbalances and skews in group composition, and their effects on curation decisions and outcomes. We also examine a type of deviant behavior where malicious users submit false opinion information in an attempt to control what fellow users see and to manipulate the choices they make.

A brief summary of the two parts of this research program and the contributions from each part is given below.

- Research Theme 1: Collaborative Curation Practices and Mechanisms
  - Analysis of curation practices and their broad effects in the English Wikipedia [89]
  - Field experiment with two different curation mechanisms on MovieLens
- Research Theme 2: Collaborative Curation Challenges
  - Analysis of the effects of group composition on curation decision quality in the English Wikipedia [87]
  - Exploration of community-scale gender imbalance and its effect on the English Wikipedia's content coverage [90]
  - Experimentation and analysis of potential attacks on the collaborative filtering algorithms that are used to address the information overload often seen in SPCs [88]

We believe that the products of this research program will benefit both SPC designers and researchers. With an improved understanding of how collaborative curation works in SPCs, we hope to inform the design of the next generation of SPCs and make them more efficient and

productive in building information repositories. We also hope that our work will provide a base of knowledge and ideas for future researchers studying the growth and evolution of SPCs.

Chapter 2

Related Work

Our research program builds on prior work on different aspects of SPCs and their use of collaborative curation to build information repositories. In this chapter, we survey the existing literature and discuss how our work extends it.

At its core, an SPC's success hinges on its ability to bring a community of volunteer users together to work collaboratively and productively to produce an information repository. This is a daunting task. Social psychology research over the past century has shown that groups of people often face process losses stemming from social loafing, interpersonal conflicts, and other coordination costs when they must work together [144]. Furthermore, simple economic analysis suggests that SPCs will suffer from the tragedy of the commons and free-rider problems, as rational users ought to have no reason to do work for an SPC without compensation [61]. In early theoretical work [154], Thorn and Connolly considered several of these issues and concluded that databases that depend on voluntarily contributed content "will be chronically undersupplied."

Indeed, not all SPCs are successful, as numerous wikis, discussion forums, open source projects, and other systems languish in a perpetual state of inactivity. Hunt and Johnson found that the distribution of download activity across open source projects hosted on Sourceforge follows a Pareto distribution, suggesting that only a few projects have substantial levels of activity, while most other projects are low-interest or dormant [76]. More recent analyses of Sourceforge-hosted projects by English and Schweik and by Wiggins and Crowston yielded

similar results [37, 166], with well over half of projects being classified as "tragedies" (abandoned projects). Arazy and Croitoru examined activity levels and patterns in over 33,000 corporate wikis within IBM, and described over 90% of them as "not very sustainable" [4]. Some SPCs are able to elicit participation from users, but still falter due to excessive spamming, in-fighting, low-quality contributions, and other unwanted and non-constructive forms of participation [141, 33].

On the other hand, many SPCs have flourished as well, resulting in vast information repositories that are curated primarily by the community. This has led researchers to seek an understanding of what motivates community members to contribute to SPCs, as well as how the repositories grew and evolved over time.

2.1 Why Collaborative Curation Works

To explore why people contribute copious amounts of time helping an SPC build information repositories without any tangible reward, some researchers have turned to social psychology research. Constructs such as the Collective Effort Model [80] and goal-setting theory [99] can offer explanations about why some people work harder than others, as well as predictions about how to increase SPC members' motivations to contribute.

Karau and Williams' Collective Effort Model describes how people decide whether to work more or less in a group than they would individually [80]. According to the model, people work harder in groups when they are striving for outcomes that they value, when they believe their contributions to the group effort will be useful and identifiable, and when they like the group they are working with. Numerous experiments have validated different parts of the Collective Effort Model on various types of online SPC information repositories including discussion forums [62, 100], movie ratings [11, 125], and movie databases [28].

Similarly, goal-setting theory suggests that people are more motivated to work when they have an explicit goal such as a target number of contributions or a specific task, as opposed to situations where people are in a self-driven environment with vague goals or no goals at all [99]. Priedhorsky et al. showed that visually highlighting specific areas that needed attention in a geographic wiki greatly increased users' likelihood to do more work [123]. In their experiments with goal-setting in a movie recommendation system, Beenen et al. also found that setting goals increased user contributions, and contrary to social loafing theory, assigning goals in the context

of a group rather than an individual increased the amount of work done per user [11]!

Together, these and other theories from social psychology, as well as ones from areas such as economics and organizational science, have framed a substantial body of work on what motivates people to collaborate with one another to curate massive repositories of information. These theories have helped researchers build a better understanding of how to build SPCs that are more effective at eliciting contributions from users.

However, motivating users to participate is only part of the challenge that SPCs face. Ensuring that the information repository is of sufficiently high quality is often an equally important concern. SPCs are often criticized for potential quality issues stemming from their reliance on collaborative curation [81, 92]. That is, because their information repositories are curated largely by ordinary users, there is no clear way to ensure a particular level of information quality. To make matters worse, pranksters, vandals, or other people with malicious intent might deliberately contribute false or nonsensical information to the repository.

Despite these concerns, researchers have found that SPCs appear to be quite capable of maintaining a good level of repository quality and organization. Giles compared the informational accuracy of Wikipedia articles and Encyclopedia Brittanica articles across several dozen topics, finding that they were roughly on par with one another [47]. Viegas et al. studied how Wikipedians decide to promote articles to the coveted "Featured Article" status and found that there was a remarkable level of formalized process and policy at work. This was especially surprising because at its core, Wikipedia is an unstructured wiki environment where one might expect more anarchy or disorganization [158].

To mitigate the problems caused by low-quality contributions, SPCs have turned to a variety of approaches. Some use automated techniques in order to reduce user burden. Sen et al. developed algorithms for identifying and filtering out low-quality tags when choosing which tags to display to describe items in an SPC [137]. In question-answering, much effort has gone into seeking ways to predict the quality of each respondent's answer, thus allowing automatic separation of high-quality answers from low-quality ones [63, 97, 138].

2.2 Collaborative Curation Mechanisms

Some SPCs include quality control measures in their collaborative curation processes, and rely on their community to be vigilant in enforcing quality norms and standards. Lampe and Resnick

studied the use of community-driven moderation of discussions on Slashdot (a technology news site) and concluded that the community does a reasonable job at identifying high and low quality comments in a fair way [91]. Wikipedians have developed specialized tools and processes for working together to fight vandals who damage encyclopedia articles [45], and have used them effectively to limit the effects of damage caused by vandals [122, 160].

Here, we note that both Slashdot and Wikipedia employ similar designs in how they allow their communities to police themselves. Both systems allow users to openly submit and publish content (comments or encyclopedia articles, respectively) without requiring prior approval, and they expect that other community members will promptly notice and respond appropriately to low-quality content. This is by no means the only possible design; SPCs have a wide variety of choices in mechanisms to influence and control how collaborative curation works.

One paradigm commonly used by social news aggregators like Digg or Reddit is social voting. In these systems, users' submissions are not immediately published to the front page, which is what most visitors read. Instead, they are sequestered into a separate less-visible area of the site and are only "promoted" to the front page if they receive sufficiently positive feedback from users who are willing to peruse and vote for (or against) their peers' submissions. Through analysis of one such site, Hogg and Lerman developed mathematical models that describe the dynamics of how users use the site, and how users' voting activity results in high-quality submissions being promoted to the front page [73].

Another design used by some SPCs is an oversight or review mechanism. That is, when a user submits content to the repository, somebody else must then review the submission and take appropriate action if it is objectionable. In two separate experiments on the movie recommendation system MovieLens, Cosley et al. studied several aspects of oversight, including who performs the review (i.e., a peer or a content expert), whether the content submitter is informed about the review, and whether a review is mandatory before publication [27, 28]. Their results indicate that the presence of oversight, regardless of who performs it, inhibits antisocial behavior, increases user motivation to contribute, and improves user perceptions of repository quality. However, oversight also comes with costs that can slow overall repository growth.

In their work, Cosley et al. also introduced and experimented with intelligent task routing (ITR), a novel technique to help potential contributors find tasks more efficiently and effectively [28]. By combining a user's previously-expressed opinions and preferences with an algorithm to filter and sort the available tasks, a site can suggest specific curation tasks that the user is

especially well-suited to perform (e.g., because they involve something relevant to the user's interests or tastes). The experiments showed that ITR was indeed effective at increasing the number of contributions and contribution reviews.

A technique that some SPCs employ to deter antisocial behavior is to introduce *barriers to entry*; that is, to require that users perform non-trivial tasks or make a payment before being given full membership privileges [5]. Of course, such an approach comes at a cost—Friedman and Resnick studied entry costs from an economic perspective and concluded that they can alleviate problems with participants repeatedly creating new accounts and behaving badly, but that they also introduce inefficiencies because the costs act as a deterrent to some would-be participants [43]. In experimentation with barriers to entry in an SPC, Drenner et al. found that subjecting users to a more arduous initiation process indeed causes some to forego participating in the SPC [36]. However, they also observed two benefits. First, users who overcame the barrier were more committed to the community, doing more and better work in the long term. Second, the design of the barrier itself was effective at shaping future user behavior.

An area closely related to collaborative curation mechanisms and SPCs is the design of human computation systems. In his seminal dissertation work on this topic [162], von Ahn defines human computation as "...a paradigm for utilizing human processing power to solve problems that computers cannot yet solve." SPCs differ in that they involve more elements of crowdsourcing and social computing, where the intent is not to have humans take the place of computers, but to harness the efforts of many people working together to collectively achieve a goal that might otherwise be intractable due to the cost of human labor. Nonetheless, human computation systems and SPCs share issues of motivating participation and maintaining quality. In [124], Quinn and Bederson present a comprehensive survey and taxonomy of the processes used in human computation systems to address these issues. There, we see a number of design elements that are also used in SPC collaborative curation mechanisms, including peer or expert review processes, reputation systems, and "wisdom of crowds" style information aggregation processes.

The design of collaborative curation mechanisms also shares some similarities to the area of mechanism design in game theory and economics [44]. In both, the mechanism designer has the goal of establishing a workflow or structure that yields desirable behavior from participants and maximizes total value gain. Because participants are assumed to be self-interested, the designer must expect that participants will behave strategically, potentially lying or exhibiting

anti-social or deviant behaviors if doing so will improve their own value gain. We note that while this parallel to mechanism design exists, we will not attempt to bridge the wide gap between theoretical mechanism design and existing practice in SPCs in this thesis.

2.3 Growth Patterns in SPCs

Unlike professionally-curated information repositories that are carefully crafted according to rigorous standards, the growth and lifecycle of an SPC can be vexing to describe. Governed by the goals and behaviors of numerous participants and their interactions with one another, collaborative curation can yield repository growth patterns that are more organic and perhaps unconventional.

The success and growth of an SPC often depends on a core set of dedicated contributors that act as leaders and exemplars for other contributors. Studies of Wikipedia have shown that a disproportionately large portion of contributions and encyclopedic value have consistently come from a small subset of its community of editors [122, 83]. Peddibhotla and Subramani's analysis of customer-contributed product reviews on Amazon.com found that the top 1,000 reviewers, who constitute just 0.08% of the reviewer community, were responsible for 7% of the reviews [116]. Despite the great quantity of reviews written by these contributors, quality did not suffer. In fact, these contributors' reviews were deemed considerably *more* helpful than those written by other, less-prolific reviewers. Furthermore, these elite reviewers often acted as early contributors for products with few (or no) existing reviews. The early reviews were especially valuable because they provided information about something for which there was previously little or no commentary, and they set the stage for others to chime in with their own reviews.

The importance of early contributors was also highlighted by Sen et al. in their exploration of the growth and evolution of an SPC's tagging vocabulary [136]. Their experiments showed that users were substantially influenced by the tags that they saw their peers using. As predicted by social learning theory, users tended to imitate or copy their peers, and as a result, the types of tags that early participants chose to add were the most influential in guiding the evolution of the tagging vocabulary.

Not all SPCs exhibit the same evolutionary patterns. In a cluster analysis of over 33,000 corporate wikis [4], Arazy and Croitoru found six clusters of wiki lifecycles that described

three broad patterns. 93% of wikis fell into clusters containing wikis that were apparently not sustainable and that became inactive soon after their inception. 6% of wikis were able to sustain moderate levels of activity for one to four years before tapering off. Finally, 1% of wikis exhibited substantial levels of activity for an extended period of time.

Suh et al. studied the growth of the English Wikipedia and showed that its growth rate has plateaued or slowed as it has aged, which bucks previous predictions of sustained geometric or exponential growth [146]. They suggested that Wikipedia's growth is best described as logistic, and proposed a growth function based on ecological predator-prey population models. The model predicts that there is upper limit of "encyclopedic" knowledge (which may change over time as more knowledge is generated), and Wikipedia's growth will rapidly decelerate as the encyclopedia approaches that limit.

2.4 Group Composition and Collaborative Curation

One often-invisible aspect of collaborative curation that can have a profound effect on the quality, quantity, and type of work that gets done is the composition of the user community. Social psychologists have long studied the effect of small group composition on performance, efficiency, conflict, conflict resolution, and so on [144, 94, 8]. Furthermore, cultural, language, and gender differences in technology use are well-known and have been observed in social tagging applications, social networks, and other technological applications (e.g. [35, 74, 75, 157]). Thus, communities of contributors that are skewed in some way may yield growth patterns that produce repositories exhibiting some form of skew or irregularity.

For instance, the effects of user geography on SPCs have been studied in some detail. In the geographic wiki Cyclopath, Panciera et al. found evidence of geographic locality in users' editing habits [115]. Users made contributions to the regions of the map that they viewed the most, which are likely to be near their home, workplace, and commuting routes. Hecht and Gergle looked at similar phenomena in Flickr and Wikipedia, finding that substantial portions of users' contributions also exhibit geographic locality [66]. In Flickr, users' photos tended to have GPS geotags indicating locations near the uploader's self-reported home location, and in Wikipedia, users tended to edit articles about locations near them.

A follow-up study by Hecht and Gergle looked at 25 different Wikipedia language editions to study the impact of primary user language on SPC topical coverage and diversity [65].

Despite all the Wikipedias sharing a common goal of building an encyclopedia containing the world's encyclopedic knowledge, the study discovered that there was surprisingly little overlap in the topical coverage between any pair of Wikipedias. Only 0.12% of topics appeared in all 25 Wikipedias, and a large majority of topics appeared in just one edition. Even when considering just the largest three Wikipedia editions (English, German, and French), just 7% of topics appeared in all three.

Likewise, when comparing the English and Polish editions of Wikipedia, Callahan and Herring found systemic differences in articles about famous Americans and Poles [20]. English-language articles and American subjects tended to be associated with democracy and personal controversies (e.g., extramarital affairs, or problems with law enforcement), while Polish-language articles and Polish subjects tended to be associated with communism, national pride, and overcoming career adversity. It appears that the primary language of a user community – and, likely, the culture that the community represents – has a profound influence on collaborative curation.

In a similar vein of research, Halavais and Lackaff analyzed the English Wikipedia's topical coverage in several different knowledge domains, and discovered substantial differences [57]. For example, one of their results indicate that Wikipedia's overlap with a physics encyclopedia is much higher than its overlap with an encyclopedia of poetry. This could perhaps be explained by Wikipedia founder Jimmy Wales' speculation that large portions of the encyclopedia's contributors are technologists from geek culture who likely have more interest in the physical sciences than in the humanities [165].

Skews in topical coverage are not the only potential effect of differences in group composition. The effect of intragroup diversity on participation and commitment in online communities has also been an active area of research. Chen et al. analyzed the effect of group tenure diversity and interest diversity on productivity and withdrawal in Wikipedia, and showed that groups that have high interest diversity and a moderate amount of interest diversity yield the highest productivity and lowest member turnover [22]. Ludford et al. found that groups comprised of people with dissimilar movie preferences participated more actively than like-minded groups in a movie discussion forum [100]. More diversity is not always associated with positive outcomes though: Lieberman et al. conducted a study looking at the effect of internet support groups for Parkinson's disease patients, and showed that compared to heterogenous groups, homogenous groups (with respect to patient age and time since diagnosis) yielded *more* commitment from

members as well as improvements in depression and disease symptoms [95]. The effect of intragroup diversity appears to be equivocal.

2.5 Deviant Behavior in Collaborative Curation

In some SPCs, there may be participants who have agendas that differ from those of their peers or of the SPC's administrators, and who may behave in ways that are not necessarily in the best interests of the overall community. For example, a company might surreptitiously market their own products in a question-answering site by submitting questions and answers that contain references to or recommendations for their products. These sorts of strategic or antisocial behaviors are an interesting consideration in the design of collaborative curation mechanisms, and go beyond the quality control concerns discussed earlier. Contributions by sufficiently clever deviant participants may be constructed in ways that "game the system" and bypass quality-control measures (e.g., the contributions may be disguised as high-quality, or designed to exploit weaknesses in automated algorithms).

Cosley et al. observed an example of such behavior in their work with oversight in SPCs [27]. Some participants ignored the provided task instructions and opted to "hijack" the task for their own purposes. The presence of oversight reduced this behavior, but did not fully eliminate it. Suh et al. propose visualization techniques that can help identify collusion and user factions in social systems such as Digg and Wikipedia [145]. Reports of similar deviant behavior and its effect on large systems have been reported by the mainstream press [107, 119, 64, 18, 60].

There is an extensive body of work exploring deviance in social computing systems related to SPCs. Peer-to-peer file sharing protocols such as BitTorrent have been scrutinized for technical weaknesses that allow dishonest participants to cheat and receive more than their fair share of resources [98, 118]. Dellarocas studied user-driven reputation systems used in online trading communities such as eBay, and formulated techniques for addressing unfair or discriminatory behavior such as colluding with customers to collectively badmouth a competitor [32].

Numerous researchers have identified and studied an apparent propensity toward deviant behaviors like deception, trolling, and flaming in computer-mediated communication systems such as synchronous chat, bulletin board systems, mailing lists, and even multi-user dungeons

(MUDs) [16, 9, 33, 34, 70]. Suler introduced the term *toxic disinhibition* to describe this phenomenon, and posited a theory explaining why online environments tend to elicit such behavior, which incorporates reasons such as anonymity, lack of status or authority cues, and escapism [147].

2.6 Contributions and Impact

The work in this thesis cuts across all of the research areas described in this chapter and builds on many of them.

Our first research theme, collaborative curation practices and mechanisms, will provide further insight on the dynamics of collaborative curation in SPCs. Past work has focused on user contribution patterns [122, 83, 116] or high-level SPC lifecycle patterns [4, 146]. We extend this work in chapter 3 by studying the processes through which items are added to the SPC repository, as well as the properties of items that are added as the SPC ages and evolves. Then, in chapter 4, which describes our experimentation with different new item curation mechanisms in an existing system, we augment existing empirical and theoretical work on curation mechanisms (e.g., [27, 28, 73]) with additional insight into how different curation mechanisms affect the trajectory of an SPC repository, and the costs and benefits associated with the different choices available.

In our second theme, collaborative curation challenges, we look beyond motivation and quality control challenges, and focus on issues of group composition and deviant behavior. Our analysis of the effects of group composition on decision quality in a real-world SPC described in chapter 5 builds on a substantial corpus of existing studies that primarily utilize controlled lab settings. Next, in chapter 6, we turn to studying the effects of large-scale gender imbalance in an SPC, augmenting existing analyses of the effects of geographic and language skews [57, 66, 65]. Finally, chapter 7 describes our pioneering work in looking at a novel type of deviant behavior targetted at recommender systems, which has helped spurred substantial follow-up work by others in the research community.

Research Theme 1: New Item Collaborative Curation Practices

Chapter 3

Collaborative Curation in Wikipedia

3.1 Introduction

In this chapter, we examine one of the world's largest SPCs, Wikipedia, looking at its collaborative curation practices and the effect they have on the encyclopedia. With millions of articles on a staggering variety of topics, Wikipedia has successfully established itself as a useful compendium of general knowledge, and is read in excess of 500 million times per day [41].

3.1.1 Related Work

Wikipedia's success has not gone unnoticed by the research community. Scholars from numerous academic fields have been intrigued by Wikipedia, recognizing it as a ecosystem consisting of many interesting processes that are ripe for study.

Community Collaboration. Wikipedia's core idea is to rely on its user community to collaboratively write and curate an encyclopedia. To this end, Wikipedia's users have the ability to edit just about anything on the site, and they collectively provide *all* of its content. At first thought, this sounds like a recipe for disaster – how can a group of ordinary people write a good encyclopedia? Why would someone volunteer to do this? What about people who do not know what they are doing? How can we know whether people are writing good articles?

Despite these challenges, Wikipedia seems to be flourishing, and researchers have worked to learn more about how and why. One analysis found that as much as 50% of the early work on Wikipedia was done by a tiny group of "elite" contributors making up less than 5% of its editor

population [83]. However, the study finds that in recent times, the balance has shifted toward a larger number of infrequent contributors, with the work done by the elites declining to less than 30%.

Bryant et al. conducted in-depth interviews with several experienced Wikipedia editors to learn more about their motivations for and experiences with "becoming Wikipedian" [17]. Their findings suggest that editors undergo a distinct transformation when moving from novice to expert, transitioning from peripheral participation to a fuller sense of community involvement where they take on more responsibilities and have richer interactions with their peers.

Conflict and Vandalism. Of course, letting anyone edit any article also has downsides. There can be disagreement among those who are working on an article, leading to conflicts and arguments about the article's content. Kittur et al. have studied ways to identify and visualize conflict, finding that it has been on the rise as Wikipedia grows [85]. Vuong et al. developed models to detect the presence of controversy by looking at how people add and delete words when editing an article [164].

Also, some users are malicious and will vandalize Wikipedia articles by deleting content, adding nonsensical content, or injecting misinformation. Viegas et al. developed a visualization tool to help understand patterns in how people's edits to an article have shaped and reshaped the article over time [160]. Using the tool, they identified and studied several edit patterns commonly used by vandals and the response they receive from the rest of the community. Their results suggest that Wikipedians are fast at repairing the damage vandals cause, with more than half of mass content deletions being addressed within three minutes [160]. Later work by Priedhorsky et al. found that the visible impact of vandalism in Wikipedia to its readers is small but rising rapidly [122].

Governance. Dealing with issues such as conflict and vandalism lead to the need for governance. To maintain order, there must be policies about what types of behavior are acceptable or unacceptable, processes to guide how conflicts are resolved, and guidelines regarding article content and style. A number of studies have focused on how Wikipedia's governance works, seeking to learn about the formation and evolution of Wikipedia's policies [40], and how those policies are applied to guide collaboration [86, 12]. Wikipedia is largely decentralized and self-governing, with many decisions, including those surrounding key policies, being made by the users themselves.

Content Quality. Finally, the content itself in Wikipedia is also of much interest. Some see

it as a vast source of general knowledge and wonder if the semi-structured content can be used in interesting ways. For instance, Milne et al. developed ways to automatically generate domain-specific thesauri from Wikipedia articles. They found that the generated thesauri offered good coverage and more contemporary language usage than expert-created ones [108]. Likewise, Wu and Weld have reported success in extracting structured information from Wikipedia, using machine learning techniques to cope with article incompleteness or inconsistencies [173].

However, because Wikipedia's content is not necessarily written by experts, much debate exists whether it is complete, accurate, and high-quality. Giles compared the quality and accuracy of Wikipedia and *Encyclopedia Britannica* and found that to the statistical limits of the study, the number and distribution of errors in the two encyclopedias was comparable [47]. Wilkinson and Huberman find that one distinguishing factor of high-quality Wikipedia articles is a larger number of editors [168], which is perhaps counterintuitive given the old adage: "too many cooks spoil the broth."

3.1.2 The Long Tail

In our work, we look at an area related to the wide breadth of article content on Wikipedia, but that has ties to the other processes described above as well: community collaboration, conflicts, and policies. Specifically, we explore Wikipedia's practices regarding new-item curation as they relate to its *long tail* of articles. We look at what the long article tail is and how it is evolving as Wikipedia's community of users curate the world's largest encyclopedia, creating, evaluating, and possibly deleting thousands of articles every day.

Before we formally state our research questions, we describe the phenomenon of the long tail and how we apply it in the context of Wikipedia. The Long Tail is a term introduced by Chris Anderson, the editor-in-chief of *Wired*, that refers to a business strategy that received much attention in recent years [3]. The long tail gets its name from the natural long-tailed and heavy-tailed distributions appearing in consumption rates of many types of consumer products such as books, songs, or movies. In these distributions, a small number of popular "hits" dominate, while a large number of items that individually have little consumption form the so-called "long tail." Many of these distributions are power laws.

The long tail strategy is to bolster a business's performance by finding ways to offer items from the long tail at little to no cost to the business. The theory is that while each long tail item only gets purchased a few times, their aggregate sales can significantly increase revenue.

Traditionally, this strategy has not been viable for physical stores because selling everything in the tail requires too much floor space. Thus, retail stores must carefully choose what items they stock, and they naturally gravitate toward high-volume "hits" from the head of the distribution.

However, the long tail strategy is usable by web-based retailers because physical storefronts are not required. Warehouse space is cheaper than retail space, so it becomes possible to stock and offer more items. Furthermore, in some domains such as digital music, the cost of storing items is negligible since the items consume little to no physical space.

The long tail has also been used to describe phenomena in non-commerce domains such as blogs, social networks, and tagging. Here, the long tail often refers to the natural long-tailed distributions found in these domains rather than to a business strategy. For instance, in their work to identify influential bloggers, Agarwal et al. describe the distribution of blog influence as a long tail [2]. In collaborative tagging systems, distributions of tag usage can be modeled using a stochastic urn process that naturally yields long-tailed distributions [51].

In the context of Wikipedia, we apply the long tail to its collection of encyclopedia articles and the viewership that each article receives. As a point of reference, consider that the 2007 edition of the *Encyclopedia Britannica* contains 65,000 articles[1], which is well less than 5% of the millions of articles in the English Wikipedia. Using a Wikipedia web log dataset that we describe in section 5.2, during the last three months of 2007 the top 65,000 Wikipedia articles ranked by visits comprise less than 60% of all visits to Wikipedia articles. So, if we consider the remainder of Wikipedia articles to be the long tail, it makes up over 40% of Wikipedia traffic, which is about 60 million article views per day as of late 2008 [153]!

Other researchers have explored different manifestations of the long tail in their studies of Wikipedia. Kittur et al's analysis suggested the emergence of a long tail of user participation in which a large percentage of work is being done by a large group of people, each of whom only does a small amount of work [83]. Wu et al. found that the use of infoboxes, a specific type of structured data found in Wikipedia articles, follows a long tail distribution, and proposed ways to improve automated information extraction methods in the presence of such a skewed distribution [172]. The present work is the first research to look deeply at the questions of the long tail in article readership, combined with an exploration of how collaborative curation influences the evolution of the long tail.

[1] http://corporate.britannica.com/library/print/eb.html

3.1.3 Research Questions

In the rest of the chapter, we first discuss the datasets we use for our analysis, and then address the following five research questions, in one section each.

RQ *Long Tail Visits*: To what extent do Wikipedia viewers look at articles in the tail?

RQ *Wikipedia Growth*: How have article birth and mortality rates changed over time as the community's curation practices have evolved?

RQ *Topic Notability*: As time passes, are the articles that survive in Wikipedia increasingly on obscure topics?

RQ *Deletion Reasons*: What are the reasons given for deleting articles? How do these reasons relate to the long tail?

RQ *Article Life Span*: When in the life of an article is it most likely to be deleted?

The present research is different from many other projects that study group dynamics in that it is a study of a single distinctive community. Because of the distinctive – some say unique – properties of Wikipedia, it is not obvious how to extrapolate these results to other communities. We argue that Wikipedia is such an important social production community, with hundreds of millions of readers every *day*, that research that helps understand how and why it works is independently interesting, even if it is not obviously generalizable. We further speculate that the successes of Wikipedia, if deeply understood, can lead to the design of computer support for other groups that can share some of those successes. However, the scope of the present work is limited to expanding our understanding of Wikipedia's curation practices.

3.2 Data

Most of our analyses are performed using the information sources that are described below.

English Wikipedia data dump files. These are several datasets that were made available on the Wikipedia database download site[2] at various times between 2006 and 2008. Each dump contains a snapshot of all articles that existed on Wikipedia when the dump was created. The data provides information about every article revision (i.e., time of creation, author, and edit comments).

In this work, our analysis is confined to the *Main* namespace. We make this distinction because we are interested specifically in the encyclopedic content available on Wikipedia, which

[2]http://download.wikimedia.org/enwiki/

is stored in the *Main* namespace. We do not consider other Wikipedia namespaces such as: *Talk*, which contains meta-discussions about articles; *User*, which contains personal information about Wikipedia users; and *Wikipedia*, which contains information specific to Wikipedia itself (e.g., help pages, community standards, content guidelines).

English Wikipedia event log. The database download site also provides a log of special events that have occurred on Wikipedia. These events include administrative meta-actions such as blocking abusive users, renaming articles, or granting users special privileges. Of particular interest to our analysis, this log also contains information about article deletions, including the reason given for deletion.

Note that by deletion, we mean the case where somebody has removed the article *and* all of its revision history from public view. We do not consider cases where a user simply erases all the text from an article, since the removed text is still readily available by browsing the article's revision history. Additional details about the article deletion process are given in section 3.6.

Sample of Wikipedia web logs. The Wikimedia Foundation has graciously supplied us with an anonymized feed of the web access log for their web servers. The feed contains the URL and timestamp of every 10th HTTP request. This log allows us to accurately estimate how many people are reading each Wikipedia article. Note that because of this sampling, all reported viewership figures are approximately a factor of ten below their actual values.

3.3 Long Tail Visits

First, we will look at the overall distribution of visits across Wikipedia articles to get a sense of what Wikipedia's long article tail looks like. For this analysis, we used a web log sample from October 1, 2007 through December 31, 2007.

To visualize this distribution, we present it as a complementary cumulative distribution function (CCDF), which is shown in figure 3.1. The value of the CCDF is equal to one minus the value of the cumulative distribution function. When plotted on a log-log scale, this representation of the data allows us to see whether the distribution is a power law by looking at whether the CCDF is a straight line. Furthermore, this representation is more robust against biases and noise in the data than the probability density function or rank-frequency plots [24, 53].

The CCDF exhibits two outliers at its far right where the number of observed visits jumps dramatically. These points are for the two articles "Main_Page" and "Wiki", which are viewed

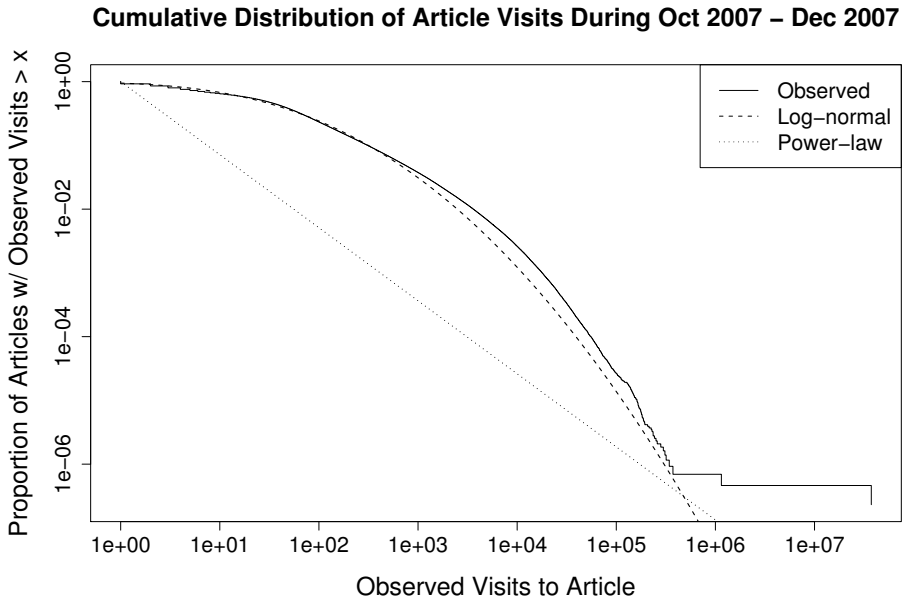

Figure 3.1: Complementary cumulative distribution function of Wikipedia article visits on log-log scales. Visits between October 1, 2007 and December 31, 2007 are counted. The dashed and dotted lines are maximum-likelihood estimate fits to the log-normal and power-law distributions.

far more often than any other article. The Main_Page article is shown to anyone who views the front page of the English Wikipedia, and is, therefore, a common landing page for bookmarks and inbound links to Wikipedia. So, its higher traffic is to be expected. The high traffic drawn by the Wiki article is more surprising. Given that Wikipedia itself is a wiki, perhaps there is a disproportionately large number of Wikipedia visitors who are curious about what a wiki is.

Figure 3.1 also shows maximum-likelihood estimate fits to the power law and log-normal distributions. The CCDF shows a distinct curve, and is a much better fit to a log-normal distribution than to a power law. This result is in apparent conflict with one Wikipedian's analysis that examines the traffic of the top 1,000 most visited articles and concludes that beyond the top few articles, the distribution looks like a power law. [3] The reason for the disagreement is that often power law and log-normal distributions appear the same when looking at only a few orders of magnitude on a log-log scale. When the Wikipedia data are extended to the full set

[3] http://en.wikipedia.org/w/index.php?oldid=222154521

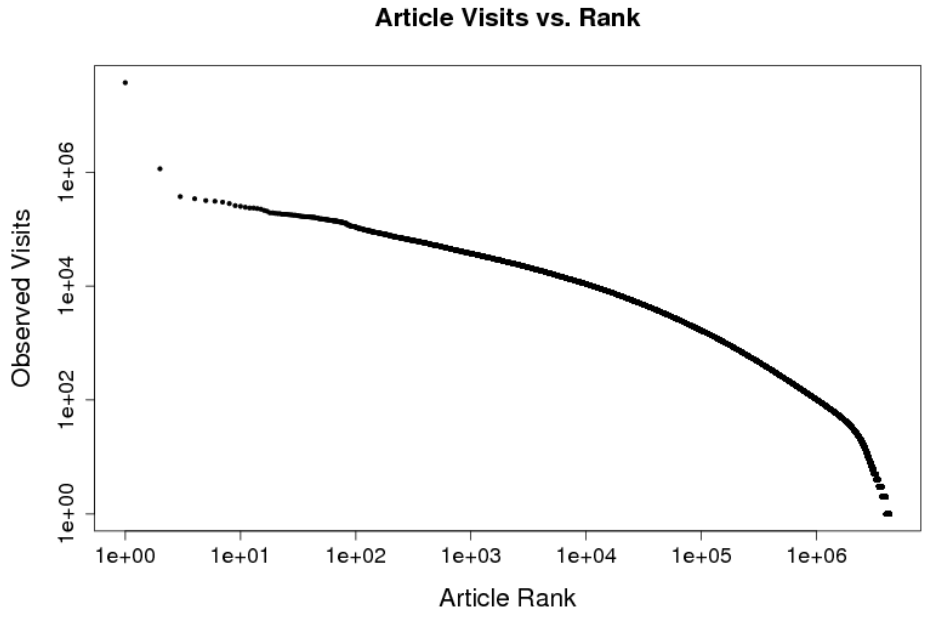

Figure 3.2: Rank-frequency plot of Wikipedia article visits on log-log scales. The top two most visited pages, which appear to be outliers, are the "Main_Page" and "Wiki" articles.

of articles, rather than just the top 1,000, the function is clearly non-linear (figure 3.2), and a power law fit can be ruled out.

The observation that Wikipedia traffic is not a power law raises interesting questions about Wikipedia's evolution. The conditions lend themselves naturally to a long tail power law scenario: practically unlimited storage for articles, low barrier to entry, and efficient digital distribution. Yet, the empirical results suggest the distribution is much closer to log-normal, which manifests itself as a truncated power-law distribution or a "drooping tail" in which there is a deficiency of low-readership articles. One possibility is that the natural distribution would be a power law, but that other factors such as efforts to deter creation of low-value articles have "truncated" the distribution such that it has become log-normal (for a discussion of the evidence on this point, see section 3.6).

Overall, in answer to **RQ *Long Tail Visits***, Wikipedia traffic does show a long-tailed distribution, especially over the first thousand articles, but it does not follow the classic power law over the entire six orders of magnitude of article popularity rank. A log-normal distribution is

a better description of the data. Some authors argue that only power law distributions should be called long-tailed, while others argue that log-normal distributions should share the name. We won't get caught up in an argument about terminology here, but will note that excluding the 65,000 most popular articles from Wikipedia – 65,000 is the number of articles in the entire Encyclopedia Britannica – still leaves 60 million article views a day for the rest of Wikipedia. So, Wikipedia traffic is substantially increased by the long tail phenomenon.

3.4 Wikipedia Growth

Before we delve deeper into Wikipedia's evolution over the years, we first look at the big picture – how has Wikipedia grown? What broad patterns have there been in the creation and deletion of articles over time?

3.4.1 Data Challenges

A major issue with the article snapshot dumps is that they do not contain any information about articles that were deleted prior to the time that the snapshot was made. The deletions are captured in the event log, but information about the deleted articles is not present. This leads to several challenges in performing analyses of article deletion behavior, since there is limited information about the vast majority of articles that have been deleted. We describe the impact of these limitations and how we worked around them in sections 3.4.2 and 3.7.1.

Also, robots and users of semi-automated editing tools[4] occasionally perform tasks that introduce substantial noise to our data. These tasks include one-off projects such as creating articles about politicians or animal species by copying information en masse from outside sources. Because these tasks undergo advance review by administrators and fellow community members[5], the created articles are typically not scrutinized, and are rarely subject to deletion. In order to focus on how Wikipedia's long article tail is affected by the curation actions of human users, we exclude articles created by known automated processes in our analyses that involve article deletion.

Finally, data about deletions occurring prior to December 2004 are unavailable, so we only focus on the period between December 2004 and December 2007 for which we have both article

[4] http://en.wikipedia.org/wiki/WP:AWB
[5] http://en.wikipedia.org/wiki/WP:BOT

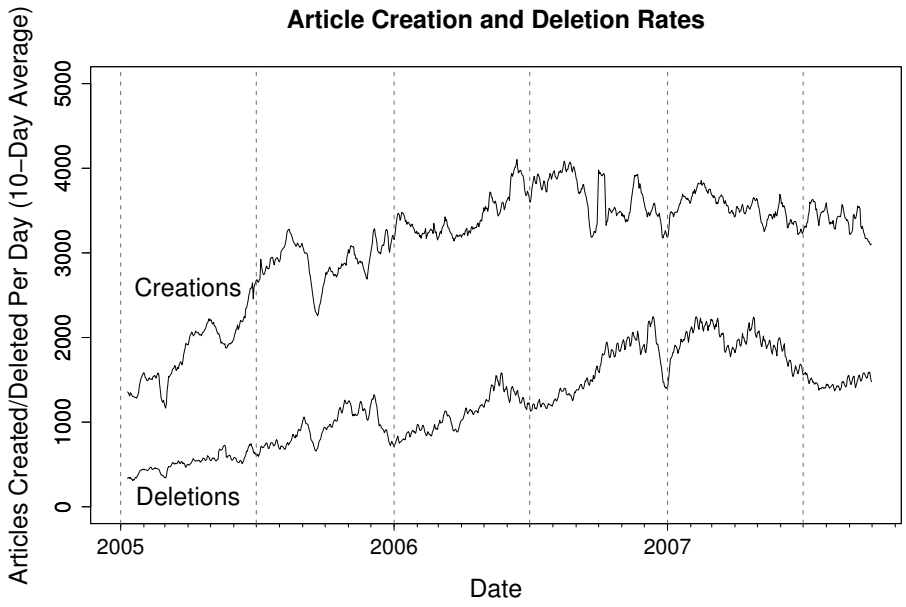

Figure 3.3: Daily rates of surviving article creation, and article deletion. Rates are smoothed using a ten-day moving average.

creation and deletion data.

3.4.2 Article Birth and Death Rates

Figure 3.3 shows the birth rate of surviving articles[6] and the death (deletion) rate of articles, expressed as a ten-day moving average to smooth out noise. We present the birth rate of surviving articles because the true article birth rate is not known due to the lack of data on deleted articles. Specifically, we do not have data that tells us when deleted articles were originally created, so we cannot determine the total number of articles created during a given interval. However, the available data do yield some interesting patterns.

Note that the death rate tends to follow the surviving birth rate, rising and falling mostly in lockstep. This correlation suggests that if an article death occurs, it tends to be near the time that the article was created. Later, in section 3.7, we use other datasets to enable careful

[6]Our article counts differ from Wikipedia's published statistics due to differences in the definition of an article. We only exclude bot-created articles, whereas Wikipedia excludes some shorter entries that they consider "non-articles," but includes bot-created articles.

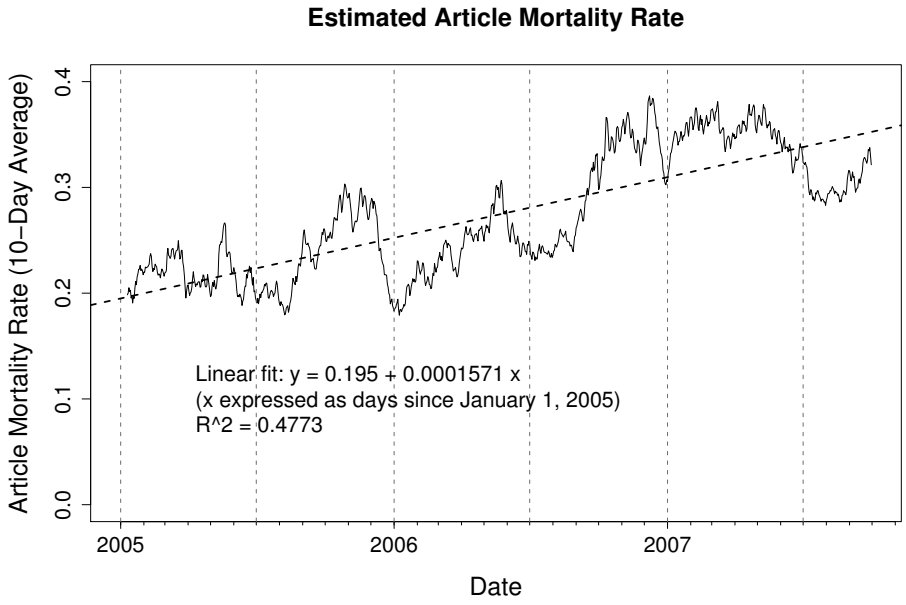

Figure 3.4: Estimated article mortality rate, smoothed using a ten-day moving average.

measurement of survival rates that allow us to validate this conjecture. For now though, we will simply assume that article deletions occur near the time of article creation. Making this assumption allows us to estimate the total article birth rate during some period by summing the death rate and the surviving birth rate.

With estimates of article birth and death rates, we can compute an estimated article mortality rate and see how it changes over time. Figure 3.4 shows this relationship. While there is much fluctuation in the mortality rate, there is a modest upward trend in mortality, suggesting that new articles are being increasingly subject to deletion as Wikipedia grows and evolves. This is consistent with results presented in Kittur, et al. that show that Wikipedians are spending an increasing amount of their efforts on indirect work – enforcing policy, dealing with vandalism, and so on [85].

We also observe that there is noticeable movement in this mortality rate that correlates to actions taken by the Wikimedia Foundation or its members. In particular, we found the following two instances.

First, in December 2005, the Wikimedia Foundation made the decision to restrict article

creation to users who have a Wikipedia account. [7] This was done in response to a high-profile vandalism incident involving an article about former *USA Today* editor John Seigenthaler, Sr. In May 2005, somebody created a hoax article about Seigenthaler that linked him to the John F. Kennedy and Robert F. Kennedy assassinations. The article was left untouched for several months before Seigenthaler learned about it from a colleague. After working with the Wikimedia Foundation to have the article removed, Seigenthaler published an op-ed article in *USA Today* describing the incident and criticizing Wikipedia. [8]

Figures 3.3 and 3.4 show that during December 2005, the article mortality rate fell by roughly 30%, while the surviving article birth rate remained unaffected. It is plausible that the new restriction dissuaded would-be vandals or pranksters from creating questionable articles such as the hoax about Seigenthaler, and therefore reduced the number of articles that required deletion. However, the reprieve was only temporary, as the mortality rate began rising again soon afterward. Perhaps the barrier of account creation was insufficient as a long-term deterrent to undesirable articles.

Second, in August 2006, Jimmy Wales, co-founder of Wikipedia, gave a keynote talk at the Wikimania conference during which he urged Wikipedia contributors to focus on article quality rather than article quantity [165]. Wales' keynote received coverage by the mainstream media, with articles appearing in the *New York Times*[9], *Wired*[10], and other outlets. Looking back at figures 3.3 and 3.4, we see that in August 2006, Wikipedia's article birth rate decelerated and the death rate accelerated, leading to a noticeably elevated article mortality rate that remained high for about ten months. It seems that Wikipedians agreed with Wales and raised the bar for what constituted an acceptable Wikipedia article.

We stress that we have no solid evidence that these actions were directly responsible for the changes observed in article curation activity. These are interesting correlations that suggest that external factors may have a profound effect on the evolution of Wikipedia.

In answer to **RQ *Wikipedia Growth***, we can say that while the number of articles in Wikipedia is growing, their mortality rate is also slowly increasing over time. Of course, from these data we cannot say whether Wikipedians are applying tougher criteria to new articles, or whether the newly created articles are less appropriate for Wikipedia. We shall return to that

[7] http://en.wikipedia.org/w/index.php?oldid=136017357
[8] http://www.usatoday.com/news/opinion/editorials/2005-11-29-wikipedia-edit_x.htm
[9] http://www.nytimes.com/2006/08/07/technology/07wiki.html
[10] http://www.wired.com/science/discoveries/news/2006/08/71535

question in the next section.

3.5 Topic Notability

3.5.1 Deletionism and Inclusionism

The observation that over one-quarter of Wikipedia articles are ultimately deleted leads us to look at a long-running conflict that has been taking place within Wikipedia's user community for years. The constant influx of thousands of articles per day is a source of concern for some editors who believe that many new articles are about topics that are too obscure and that are not interesting enough to warrant the existence of a Wikipedia entry. These editors see such articles as diluting the overall value and credibility of Wikipedia. In contrast, other editors believe that the ever-growing set of articles is beneficial since it opens more opportunities for people to participate, and emphasizes Wikipedia's strengths as a digital resource that has no practical limit on size.

These two philosophies have been labelled as *deletionism* and *inclusionism*, respectively, and the results of their influences on Wikipedia and its long tail will be the primary focus of the remainder of this chapter.

We now look more closely at how Wikipedia and its long article tail have evolved over time. How do articles that were created years ago compare to more recently created articles? How true are deletionist concerns that Wikipedia's newer articles are increasingly about obscure topics? What did Jimmy Wales see that triggered his call for focusing on quality rather than quantity?

3.5.2 Data Challenge: Notability

To approach these questions, we need a way to measure the relative obscurity or popularity of an article. For this, we first turn to Wikipedia's standards regarding this issue. The English language Wikipedia's community has established a basic criterion called *notability* to determine whether a particular topic is worthy of an article. There are a wide range of opinions on the definition of notability and how much it should be taken into account when deciding whether an article belongs in Wikipedia. Much debate between inclusionists and deletionists has taken place on Wikipedia regarding notability, and the notability guidelines are often invoked when discussing whether to keep or delete an article that has come under scrutiny.

To pass Wikipedia's general notability guideline[11] as of late 2008, an article's topic must have "received significant coverage in reliable sources that are independent of the subject". Wikipedians have also established additional domain-specific notability guidelines for things such as books, films, and numbers[12]. These guidelines, while well-articulated, are often imprecise and open to interpretation (e.g., what exactly constitutes "significant coverage"?).

Thus, in this research, we do not propose a way to directly operationalize notability. Instead, we will use metrics that measure *popularity*, which is a related notion that may correlate well with notability in practice. While it is true that popularity is not exactly the same as notability, and that the metrics we use are unreliable in individual cases, we believe that our metrics are a good proxy for notability if taken in aggregate.

3.5.3 Readership

The first metric we will consider is readership. We measure this by counting the number of visits to each article as given in the Wikipedia web log sample, again using the interval October 1, 2007 through December 31, 2007. Articles that are read more frequently are presumed to be about things that are more well-known and interesting to Wikipedia readers, so this metric estimates how popular or obscure an article's subject is.

Figure 3.5 shows the average readership of Wikipedia articles as a function of when the articles were created. There is a striking downward trend indicating that newer articles are being viewed far less frequently on average than older ones. This suggests that newer articles tend to be about topics that draw less interest from Wikipedia readers, and are thus more likely to be in the long tail.

One confound here is that newer articles are disadvantaged because they have had less time to integrate themselves into the link structure of Wikipedia, and thus, have fewer backlinks (i.e., other Wikipedia pages linking to them). This deficiency of backlinks may result in newer articles receiving less traffic since users browsing Wikipedia encounter fewer links to new articles than to old ones. In turn, one might surmise that traffic to new articles will start low and accumulate over time as more links are created

To investigate, we control for the possible backlink effect by repeating our analysis but

[11] http://en.wikipedia.org/wiki/WP:N

[12] This particular guideline was, in part, prompted by a deletion debate over an article about the number 3.14, a common approximation of the mathematical constant Pi.

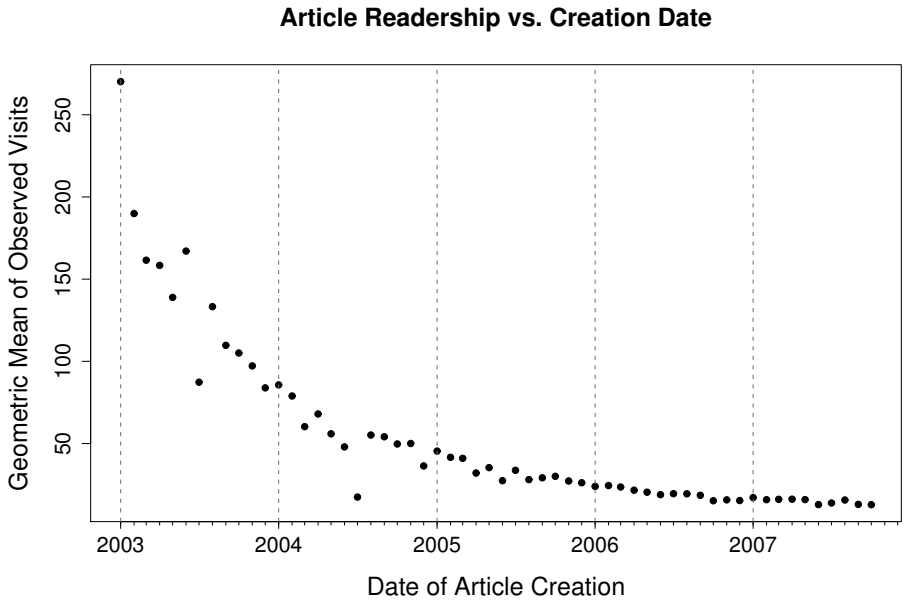

Figure 3.5: Geometric mean of the readership of articles plotted by month of article creation.

grouping sets of articles that have similar numbers of backlinks. Figure 3.6 shows that a similar downward trend still holds, although the drop over time is smaller, and now appears more linear for each group. Apparently there is an important effect of number of backlinks in explaining article traffic, but it alone does not fully explain the readership differences between older articles and newer articles.

Additionally, we did an analysis to quantify how article readership changes over time. (Articles may gain readership because of increases in backlinks within Wikipedia, because of links from the Web as a whole, because of improving position in search engines, etc.) We hypothesize an asymptotic effect exists where articles gain readership for some time before approaching a stable state that represents its "true" popularity. Thus, over time, newer articles should gain readership while older articles remain stable (effectively losing readership relative to the whole population).

To test our hypothesis, we compared the readership figures described above with figures from July 1, 2008 through September 30, 2008 to see what had changed after six months. Figure 3.7 shows the relative change in article readership share between the two data sets, again as a

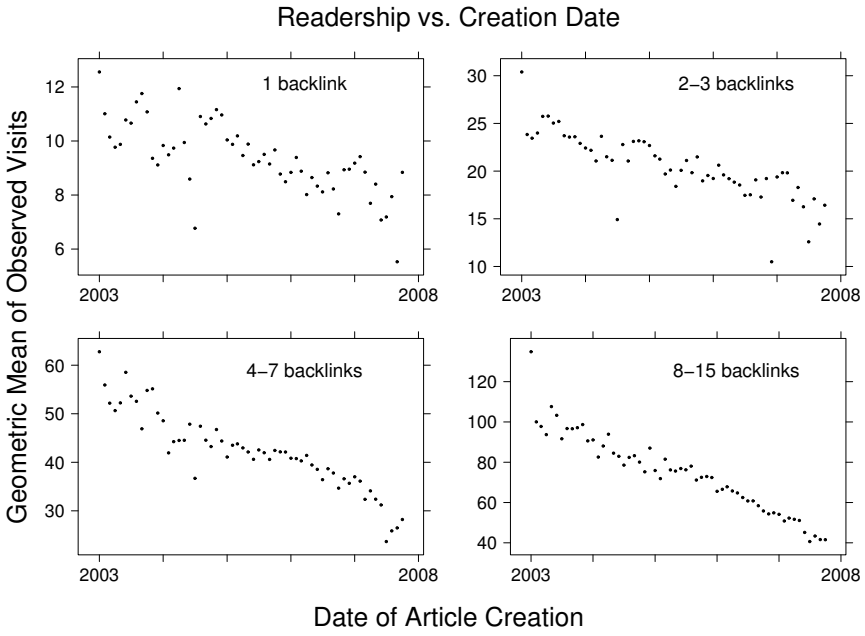

Figure 3.6: Geometric mean of the readership of articles plotted by month of article creation, grouped by articles with similar numbers of backlinks. Plots of articles with 1 backlink, 2-3 backlinks, 4-7 backlinks, and 8-15 backlinks are shown.

function of article creation date. We see that contrary to expectation, older articles increased their readership share at the expense of newer articles, which actually lost readership share as they aged!

One possible explanation for this is that newer articles tend to be about things that naturally have initial bursts of interest, such as current events and new movies or video games. Overall interest in such topics then declines over a period of time before reaching some stable state. Further research is needed to confirm or refute this explanation, but in any case, our results do not show evidence that measuring popularity using the readership metric is biased against newer articles.

3.5.4 The Search Engine Test

A second metric that we use to approximate notability is the search engine test. This is also known on Wikipedia as the *Google Test*[13]. This metric is defined as the number of results that

[13]http://en.wikipedia.org/wiki/WP:Google_test

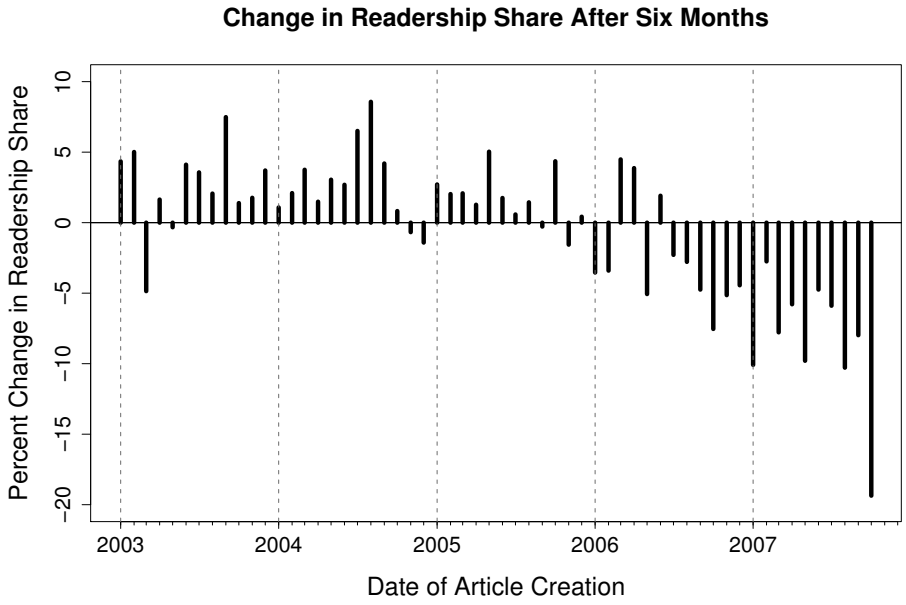

Figure 3.7: Relative change in total readership share between October-December 2007 and July-September 2008, plotted by month of article creation. Only articles created between January 2003 and October 2007 are considered.

a search engine returns when queried for web pages about a particular topic. The search engine test provides an estimate of popularity that has the advantage of being mostly independent of Wikipedia. The presence of Wikipedia and sites that copy its content affect the values, but their effect likely affects all topics similarly, and the effect is probably small compared to the size of the web.

However, Wikipedia's article about the search engine test gives several caveats in using it to establish the popularity or notability of a topic, and states that the test's result alone should not be considered to be authoritative. One major issue described is that "search engines do not disambiguate, and tend to match partial searches." The Wikipedia discussion provides a simple example: the Renaissance painting *Madonna of the Rocks*. Depending on how a search engine query is formulated, there might be many search results about the pop singer *Madonna*, which would inappropriately make it appear as if this painting was much more popular than other well-known Renaissance paintings.

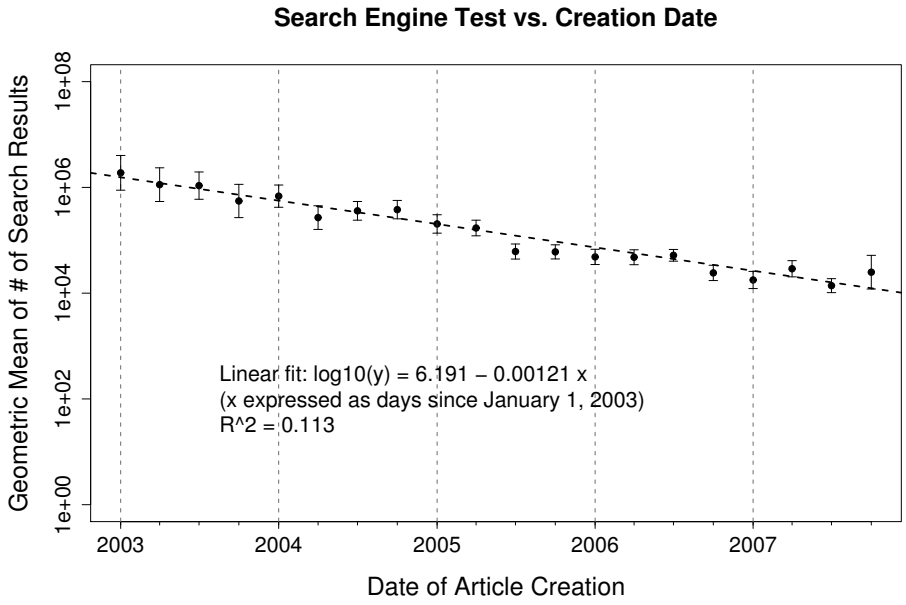

Figure 3.8: Geometric mean of results of Search Engine Test plotted by month of article creation. Geometric standard error bars and a best-fit line computed from unaggregated log-transformed data are plotted as well.

We attempt to control for this problem by restricting our analysis to articles that have single-word titles. While there is still opportunity for ambiguity (e.g., *jaguar* could refer to an animal, a car manufacturer, or a football player), we believe it reduces the effects of the problem sufficiently for our purposes. Also, using single-word titles eliminates the challenge of formulating queries for multi-word titles (i.e., word order or use of quotes), as well as confounds arising from differences in the distribution of the number of search results for multi-word searches versus that for single-word searches.

We chose a random sample of 5,758 articles with single-word titles and issued basic queries against the Yahoo! search engine using their API. [14] We were unable to test all single-word title articles in a reasonable amount of time due to limitations imposed by Yahoo!'s usage policy. Figure 3.8 shows the relationship between the mean number of search engine results and the Wikipedia article creation date. We see a downward trend similar to the one shown previously

[14] At the time that this study was done, Yahoo! was the only major search engine offering a reasonable API and terms of use for programmatically issuing numerous search queries.

for article readership, thus reinforcing the support for the results obtained using the readership metric: newer articles tend to be more concentrated in the long tail and are effectively lengthening it.

The readership data and the search engine test both provide the same answer to **RQ *Topic Notability***. The articles being created in Wikipedia *are* increasingly obscure as time passes, and are thus likely to be less notable. We do note that these data alone do not resolve the debate between inclusionists and deletionists. After all, the long tail of not-so-popular articles is responsible for a substantial number of Wikipedia page views.

However, the data might provide a principled way to reason about the cost versus value of adding articles to Wikipedia. For instance, this question could be put on an economic footing by valuing article readership in dollars, and by estimating the cost of the resources required to maintain each article. Ones attracting insufficient interest to justify their cost would be deleted. (Economic motivations are not the only way to select articles that belong in an encyclopedia; this is just one possible way to frame the debate.)

3.6 Deletion Reasons

The deletionists have likely seen evidence of these notability trends, and argue that Wikipedia is increasingly becoming a haven for irrelevant material that should have failed the test for notability. Some deletionists are working hard to seek out articles they feel are not notable, and to remove them from Wikipedia. In this section we study their success at this task, looking at the frequency of deletes, the reasons for deletes, and the changes across time in these characteristics. We are particularly interested in the effect these changes are having on the evolution of the long tail in Wikipedia.

Wikipedians have established several different processes for deleting articles. [15] We note that unlike most of Wikipedia's editing actions, only privileged users who have been given administrative privileges are allowed to actually delete an article. However, anybody is allowed to request and discuss article deletion. Wikipedia's deletion processes as of early 2009 are summarized as follows. [16]

[15] http://en.wikipedia.org/wiki/WP:DP

[16] In early 2010, Wikipedians introduced a new deletion process for addressing biographies of living persons that do not provide adequate sources; the analyses that we do predate the introduction this process, and thus are not affected by it.

Criteria for Speedy Deletion. This is the most lightweight process for deletion. There are several dozen reasons for which an article can be deleted without requiring a discussion. Among these are vandalism, advertising, or insufficient content. This process is intended to be used for uncontroversial deletions.

Proposed Deletion (PROD). This process is used when somebody believes that an article should be deleted, but for a reason not covered by the Criteria for Speedy Deletion. If no one objects to the proposed deletion, then the article is deleted. If there is an objection, the issue is escalated to the Articles for Deletion process.

Articles for Deletion (AFD). In this process, interested members of the community examine the article under scrutiny and discuss what should happen to it. Discussions last at least five days, after which time an administrator reviews the debate and takes appropriate action. This process was previously also known as "Votes for Deletion" (VFD), but was renamed because the goal is to make decisions based on community discourse rather than majority vote.

To analyze why articles are deleted, we use the event log dataset, which includes the comment left by the deleter for each deletion event. The comment is intended to convey the reason that the article was deleted. By analyzing these comments, we can gain insight into why over one-quarter of all created articles are deleted.

We scanned deletion comments for key words or phrases that refer to Wikipedia's article deletion policies. For example, the deletion comment for an article that was deleted via the Proposed Deletion process typically contains a link to the Wikipedia policy page that describes the process, *WP:PROD*. Thus, we can identify such deletions by looking for a *WP:PROD* link. We also looked for other textual indicators of this process, such as "proded", "prodded", and "proposed deletion". We created similar lists of key words for identifying other reasons for deletion. Approximately 85% of the deletions studied could be categorized in this way.

In total, we looked at 1,567,543 deletion comments for deletions occurring between December 2004 and March 2008. Using our approach, we classified deletions into seven broad classes, which are summarized in table 3.6.

Figure 3.9 shows the overall frequency that each of these classes of deletion reasons was observed in the deletion comments.

Only 12% of deletions go through the more heavyweight processses (Proposed Deletion, Articles for Deletion, or Votes for Deletion). A large majority of deletions are considered uncontroversial and are covered by the Criteria for Speedy Deletion. We see that the most

Class	Deletion Reasons
Inappropriate Content	Patent nonsense; vandalism; attack pages; blatant advertising; copyright infringement
No Content/Context	Insufficient context to identify subject of article; insufficient substantive content
Notability/Significance	Failure to assert importance or significance; non-notable subject
PROD/AFD/VFD	Proposed deletion; articles for deletion; votes for deletion
Wiki Maintenance	Redirect to a non-existent page; technical deletion (used for renaming or moving articles, merging article histories, and other maintenance-related tasks)
Other	Creator requests deletion; creation of previously deleted material; all other policies
Unknown	No recognized key words or key phrases

Table 3.1: Classes of deletion reasons.

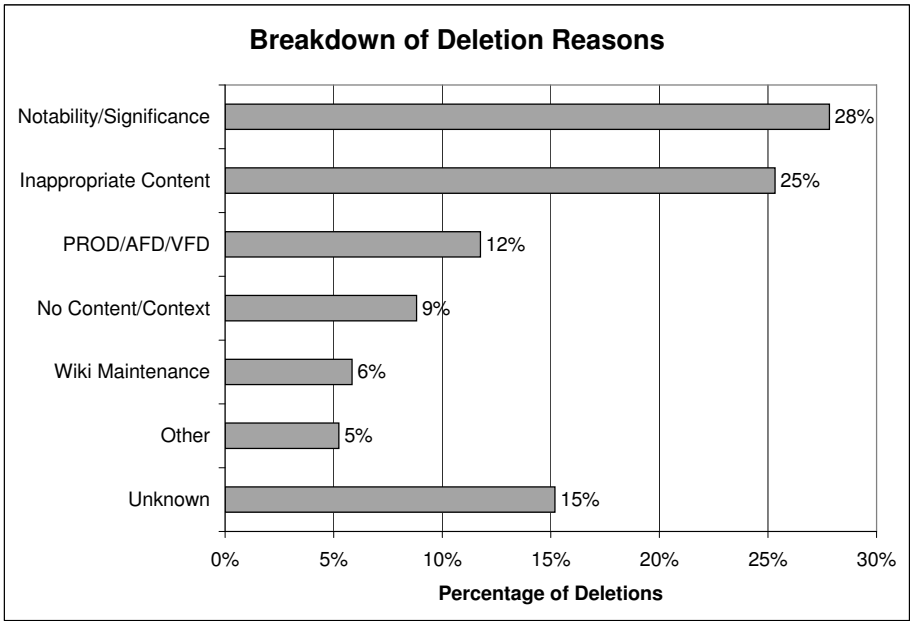

Figure 3.9: Overall frequency of classes of reasons given for Wikipedia article deletions. PROD/AFD/VFD denotes deletions occurring as a result of the Proposed Deletion or Articles for Deletion processes.

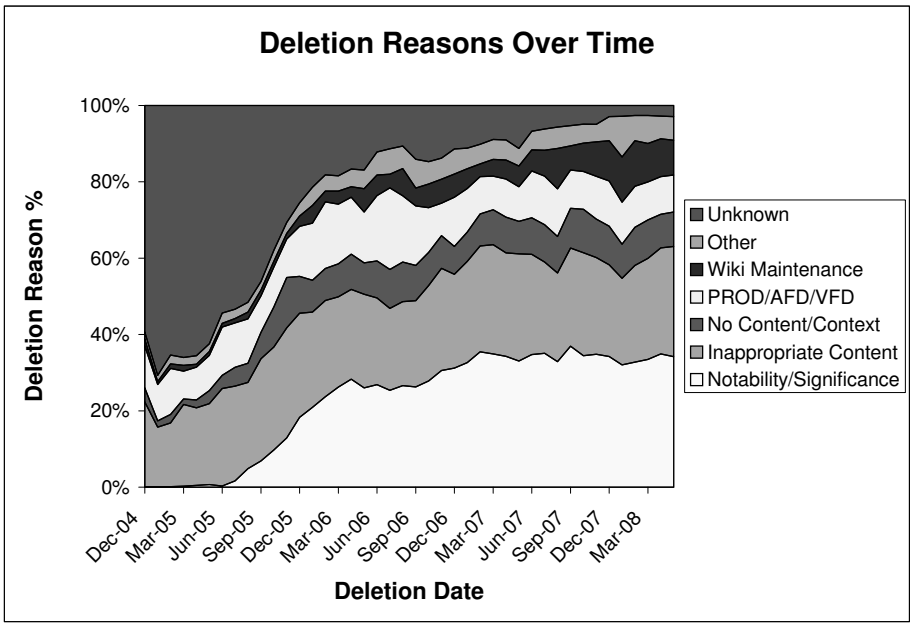

Figure 3.10: Frequency of classes of reasons given for Wikipedia article deletions by month. PROD/AFD/VFD denotes deletions occurring as a result of the Proposed Deletion or Articles for Deletion processes.

frequently-cited reasons for deleting an article are notability-related, making up over a quarter of all deletions. Next, deletions due to inappropriate content (25%) or insufficient content (9%) together make up just over a third of article deletions. Wiki Maintenance and Other are both around 5% each. Finally, 15% of deletions, labelled "Unknown" in the figure, could not be categorized using simple keyword analysis.

Figure 3.10 shows the relative frequency of deletion reasons across time. We see two noteworthy trends here.

First, the proportion of unknown deletion reasons is declining, which means an increasing proportion of deletions are accompanied by recognizable citations to Wikipedia policy. This trend is consistent with the findings in Beschastnikh, et al. that show an temporal increase in policy citations during Wikipedia discussions [12].

Second, the proportion of deletions due to reasons classified as Notability/Significance has increased over time. As we saw in section 3.5 (figures 3.5 and 3.6), there has been a lengthening of the long tail as article creators push the boundaries for what is considered notable. Our

observations here suggest that some in the community are pushing back, actively scrutinizing articles and deleting those that are deemed not notable enough.

One thing that we cannot tell from studying deletion reasons, however, is whether interpretation and application of the notability guidelines has been consistent. Are the articles that are being deleted *actually* less notable than the articles that survive? One way to approach to this question is to apply our notability proxy metrics to articles that have been deleted due to lack of notability.

The readership metric is difficult to use here, because as we will see in section 3.7, the lifetime of an article before it is deleted is usually too short to gather meaningful data. We can easily apply the search engine test though, as it does not depend on Wikipedia-specific data. On a random sample of 959 articles with single-word titles that were deleted due to lack of notability, the geometric mean of the number of search results is 6,832. This is below the average number of search result for surviving articles in Wikipedia, which, according to figure 3.8, is well over 10,000, even for the most recently created articles. The comparison suggests that the deletion decisions being made regarding notability are generally consistent with the search engine test.

However, we note that if the downward trajectory seen in figure 3.8 continues at its historic pace, then articles created in mid-2008 will have an average number of search results of around 6,000, which is comparable to that of articles that have been deleted in the past for lack of notability! Over the long term, the declining notability of new articles will lead to one of two possible outcomes. An inclusionist might hope that notability standards will become less stringent. On the other hand, a deletionist might hope that notability criteria will remain stable, and that a higher percentage of newly created articles will be deleted.

These data provide a mixed answer to **RQ *Deletion Reasons***. Overall, the "lack of notability" reason has dramatically increased in usage between 2005 and the present. However, its increase has been very slow since early 2006, and nonexistent since early 2007. The distribution of reasons given for article deletions appears to have reached a steady state. Also, deletion decisions seem to be consistent with the search engine test for topic notability.

3.7 Article Life Span

Finally, we explore the life span of Wikipedia articles and look at *when* articles get deleted during their lifetimes. How quickly does the community scrutinize new articles and make decisions about them? Was the Seigenthaler incident the norm or the exception? Are deletionists trimming the long tail, or is it here to stay?

3.7.1 Data Challenge: Article Creation Dates

Recall from our discussion in 3.4 that the data dumps are deficient in that they do not include detailed information about most deleted articles. In particular, we lack the creation date of many deleted articles, which makes it difficult to generally determine a deleted article's life span. We address this limitation of our datasets in three ways:

Direct Data Analysis. First, we directly use the Wikipedia dumps to obtain what information we can about article life span. By combining an older snapshot of Wikipedia articles with a newer event log, we can see which of the older articles have been deleted after the time that the snapshot was taken. This gives us life span information about long-lived articles, but only provides limited and flawed information about short-lived articles.

To illustrate this issue, suppose that we want to learn about articles with a life span of less than 2 days. The only articles we could examine are those that were created during the 2 days immediately preceding the time that the article snapshot was taken. If an article was created before this interval and was deleted within 2 days, then it would not have appeared in the snapshot. To make matters worse, there is an additional confound: articles that were created *and deleted* during the window of interest would also be absent from the snapshot and missed by the analysis, which leads to an undercount of articles with a life span of less than 2 days.

Inference-Based Analysis. To help augment our knowledge about very short-lived articles, we use an inference-based approach using the article snapshots and event logs. Consider a snapshot taken at time t containing the set of all articles A existing at time t, and an article deletion event for some article a that occurs at time $u = t + 1$ day. If $a \in A$, then we know the creation date of a, and can use the log analysis approach previously described.

On the other hand, suppose $a \notin A$. Then we do not know the creation date of a. However, we do know that a must have been created after time t, since by definition A contains all articles that existed at time t. We also know that a must have been created before time u because an

article cannot be deleted before it is created. Therefore, despite not knowing its exact creation time, we can still infer that a's life span is less than 1 day.

Applying this logic to all articles deleted in the first n hours after an article snapshot allows us to count how often articles were created and deleted during that n hour interval. This provides a basis for making estimates about articles that have very short life spans. However, this approach can be used just once for each article snapshot, and only provides information about articles over a small slice of time. It is, therefore, subject to the same issues that affect small sample sizes – high variability and questionable precision.

Near Real-Time Observation. To help solidify our data about short-lived articles, we turned to the Wikipedia API[17], which can be queried for information about article creations that occurred during a given interval. This data is subject to the same shortcomings as the data dumps: articles that have been deleted do not appear in article creation listings. However, the adverse effects of missing data can be greatly reduced by issuing API queries often, thus capturing article creation events in near real-time. For our analyses, we collected article creations every five minutes over a two week interval in September 2008. We then used an event log from October 2008 to determine whether the created articles had been deleted, and if so, when.

3.7.2 Life Span Results

Combining all the amassed information, we found that most articles have either a very short life or a very long life. If an article is deleted, then the deletion usually occurs very early in the article's life, quite often within the first few days. Recall that in section 3.4, we conjectured that if an article is to be deleted, then the deletion will occur near the time that the article was created. Here, we will present data that supports this supposition.

Analysis of our inference and real-time observation data shows that for any given 24-hour period, about 61% of deletions during the period are targeted at articles that were created during that period. This allows us to generate estimated survival curves with our article snapshots. The bottom-most line plotted in figure 3.11 shows our estimated survival curve for articles created during the last 24 hours before the November 2006 article snapshot. The first-day deaths are estimated, but the remainder of the curve is actual data. Interestingly, over 20% of articles survive less than a day, and about 25% survive less than two weeks. Beyond that, just another 5% of articles are deleted over the following two years.

[17] http://www.mediawiki.org/wiki/API

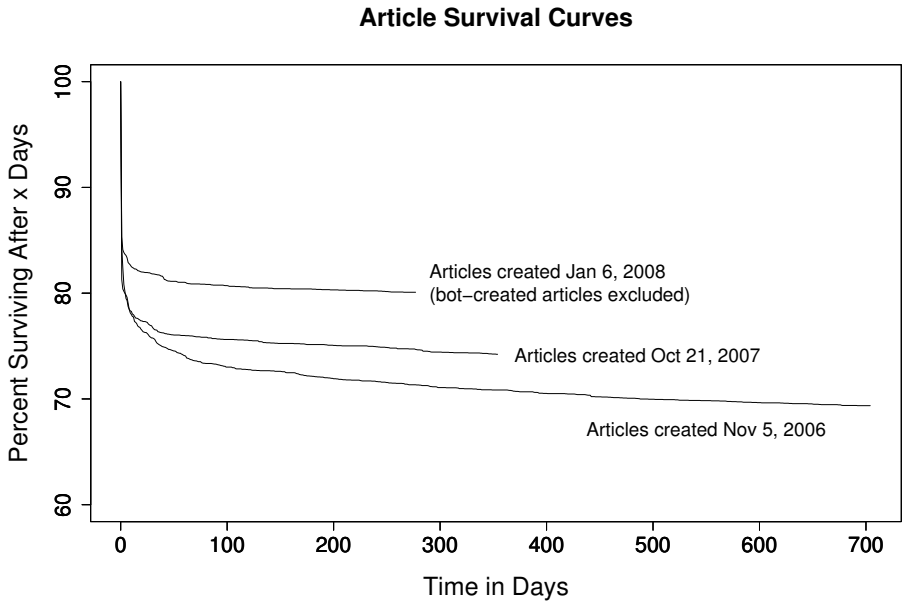

Figure 3.11: Survival curves of Wikipedia articles created during three 24-hour spans. The first-day death rates are estimated as described in section 3.7, while all remaining data is observed.

Figure 3.11 also shows survival curves generated similarly from October 2007 and January 2008 snapshots. The shapes of the curves are similar, although they "flatten out" at different percentage levels. This reflects the volatile mortality rates shown previously in figure 3.4. In all three survival curves, we see that a large majority of deaths occur during the first few days of an article's life. Wikipedians make inclusion and deletion judgments about articles very quickly, and it is uncommon for the community to return to articles later and delete them.

We also examined the question of whether deletions are "persistent" – that is, if an article is deleted, does it stay deleted, or does someone create the article again later? To measure deletion persistence, we compared our 2006 and 2008 article snapshots and looked at which of the articles existing in 2006 had been deleted in the interval between the snapshots. Of the deleted articles, we looked at what proportion of them exist in the 2008 snapshot to determine whether the deletion was persistent.

The results of this analysis are shown in figure 3.12, which shows the proportion of deletions that are persistent as a function of the article's age at the time it was deleted. We see a trend that shows deletions occurring early in an article's life are more likely to be persistent than

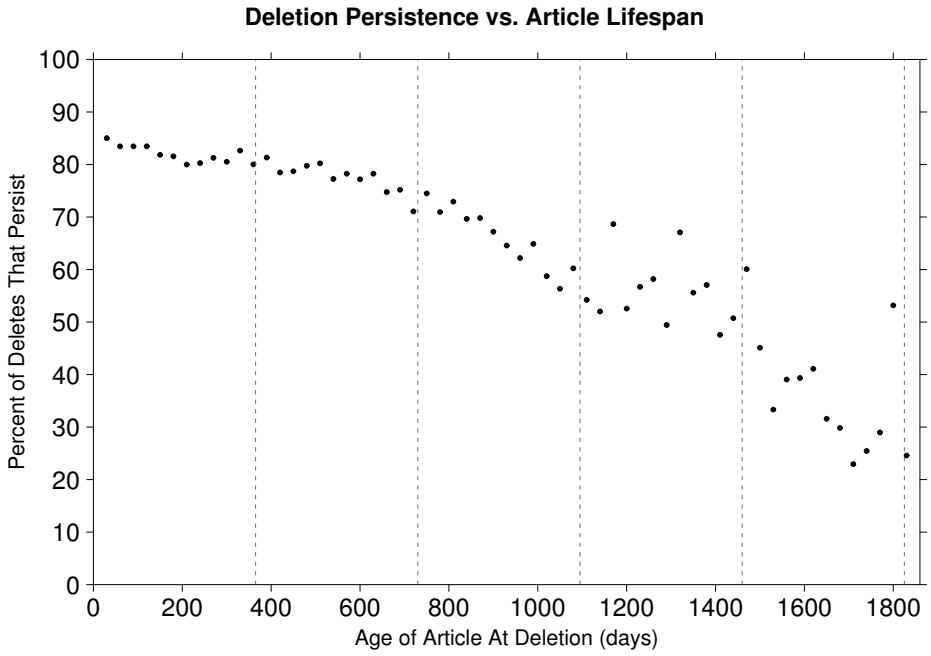

Figure 3.12: Persistence of deletions of Wikipedia articles, plotted by age of article at deletion.

deletions that occur later in an article's life. So, not only are articles unlikely to be deleted late in their life, but if a deletion does occur, it is less likely to be persistent. A common reason that a deletion is non-persistent is that the deletion was done for maintenance reasons that are tangential to whether the article is appropriate for Wikipedia. For example, an article might be deleted if a related article is being renamed to replace it.

These observations lead us to an answer to **RQ *Article Life Span***. Wikipedia's articles are here to stay, including those in its long tail. Once an article has survived the first few days of life, the chance that it is persistently deleted at some later date is small.

3.8 Discussion

In each of the preceeding sections we gave a nuanced answer to one of the five research questions. Here we briefly summarize those questions and answers.

RQ *Long Tail Visits*: To what extent do Wikipedia viewers look at articles in the tail?

The visit distribution to articles in Wikipedia follows a log-normal curve. The top articles

are by far the most popular, but the long tail accounts for a substantial fraction of visits to Wikipedia.

RQ *Wikipedia Growth*: How have article birth and mortality rates changed over time?

Wikipedia's article count continues to grow by thousands of articles per day. However, the birth rate is steady and the article mortality rate is slowly increasing, suggesting that the rate of growth has peaked and may begin declining.

RQ *Topic Notability*: As time passes, are the articles that survive in Wikipedia increasingly on obscure topics?

Yes. New articles that are added to Wikipedia *are* increasingly on obscure topics as measured by our readership and search engine test metrics.

RQ *Deletion Reasons*: What are the reasons given for deleting articles? How do these reasons relate to the long tail?

The most common reason for deleting articles is "lack of notability". The use of the notability argument is evidence of resistance within the community to including articles that are arbitrarily far down the long tail of potential Wikipedia subjects.

RQ *Article Life Span*: When in the life of an article is it most likely to be deleted?

Most articles either have a very short life or a very long life. There is little evidence to date that the long tail is effectively being trimmed over time.

Analysis alone cannot resolve the debate about whether the diversity of "long tail" articles strengthens Wikipedia, or whether these obscure articles weaken its encyclopedic nature. This debate is over what determines the health of an online user-maintained encyclopedia. Since such encyclopedias have only existed for several years, it is no surprise that there is as of yet no clear answer.

Analysis can, however, help frame the debate. For instance, it is interesting that the probability of a new article being deleted has been increasing steadily over the past three years. The articles deleted for "lack of notability" that were analyzed using Yahoo! Search for this study had average estimated notability less than that of surviving articles. However, since the estimated notability of newly created articles that survive has been declining in recent years, Wikipedia seems to have reached an intriguing inflection point: the articles that survive may be of comparable notability to those that are deleted. How will the conflict be resolved?

Chapter 4

Collaborative Curation in MovieLens

4.1 Introduction

In this chapter, we move from offline exploration of Wikipedia's curation practices to online experimentation with different curation mechanisms. We explore the design of a crucial curation element of an SPC: a mechanism for 1) community members to propose new items to be added to the repository, and 2) the community to evaluate whether proposed items are appropriate for inclusion.

We note that not every SPC requires selectivity or community involvement in new-item curation practices. For instance, Flickr and YouTube impose very few content quality guidelines, and thus, they use a simple mechanism that allows users to submit photos or videos freely without a selection or approval mechanism. However, for SPCs that desire more quality control or that would like to be selective about what items are in the repository (e.g., for ideological, technical, or moral reasons), the design of this mechanism can have a large factor in shaping the growth and direction of the repository.

On one hand, a design with insufficient community oversight might result in the overall quality of the system declining due to many irrelevant and low-quality items being added to it. For example, a common complaint with Usenet (an open and distributed Internet discussion system) is a poor signal-to-noise ratio, which can make it difficult to derive value from participating in Usenet discussions [129, 159]. On the other hand, introducing too much formal oversight to a design may result in a mechanism that is considered overly complicated, heavyweight, or bureaucratic, leading to user dissatisfaction and a decrease in motivation to contribute.

As a concrete example of the potential impact of mechanism design, consider Nupedia and Wikipedia, two ambitious projects seeking to produce online encyclopedias written and maintained by communities of volunteer editors [132]. Jimmy Wales and Larry Sanger co-founded both systems just months apart in the early 2000s. In Nupedia's editorial model, domain experts wrote and vetted articles through an extensive seven-step peer review process. In contrast, Wikipedia invited the general public to write and edit articles, and offered no formal review process to help ensure quality or correctness.

Nupedia and its intricately designed processes ultimately faltered, producing just two dozen "completed" encyclopedia articles in the three years it was online. In comparison, as seen in chapter 3, Wikipedia succeeded beyond all expectations. Even years after its inception, its editors continue to create tens of thousands of articles every month.

Why did these two endeavors, each with the same leadership, stakeholders, and goals, yield such different communities and outcomes? One of the key differences between the two encyclopedias was in their respective mechanisms for item (article) creation. While well-intentioned, Nupedia's intricate peer review process was apparently too much of an impediment to attract a critical mass of participants.

A useful theory for reasoning about different mechanisms and how they might affect motivations, behaviors, and outcomes is Karau and Williams' collective effort model (CEM) [80]. The CEM describes several factors in group collaboration and how they predict that people often work harder individually than in groups (i.e., why social loafing or free riding occurs). Some of these factors include participants not believing their effort is important to the group, not perceiving sufficient value in the group's collective work, and disliking the group.

In our work, we apply the CEM to study and compare mechanisms modeled after the ones used in two successful systems: Wikipedia and Reddit. These mechanisms differ primarily in two ways: 1) when the community scrutinizes proposed items, and 2) what level of approval is required from the community in order for an item to become available in the system's primary information repository.

Wikipedia's *wiki process* allows any registered user to write and publish an article about any arbitrary topic. New articles are immediately "live" and are part of the encyclopedia. Community review of new articles occurs on an ad-hoc basis after the articles are made available. Low-quality articles may be improved by the community of editors, or considered for deletion via one of several community-driven processes (see section 3.6 for a detailed discussion).

On the other hand, the social news aggregator Reddit uses a *social voting process*. Any registered user can submit a link (URL) to some piece of content that he or she considers interesting. However, newly submitted links are segregated into a "new links" section. Users may browse these links and cast votes for or against each one to express like or dislike for it. Links that receive a sufficient influx of net-positive votes are *promoted* to Reddit's front page, which many people use as a source for interesting content.

The full design space of mechanisms and processes for community-maintained sites is large and beyond the scope of this work. A wide variety of research topics spanning several disciplines play a role in mechanism design, including social choice theory (i.e., voting systems), crowd psychology, and game theory. In our experimentation, we seek to explore and understand the tradeoffs between two simple but quite different mechanisms. We hope that our findings can guide the analysis and design of other, perhaps more complicated, mechanisms.

In this work, we compare the wiki and social voting processes by conducting experiments in the movie recommender system MovieLens. MovieLens is a free and publicly available service used by thousands of users a month, and has been a research testbed for dozens of studies spanning numerous fields and disciplines including collaborative filtering [102, 68], intelligent user interface design [136, 161], social psychology [11, 128], and mobile technologies [104].

4.2 Hypotheses

One of the CEM's overarching predictions is that amount of effort that a person exerts in a group task is related to whether they think their effort will lead to valuable outcomes when combined with the group's efforts. Here, the value of an outcome is measured from the perspective of each group member, not from the community or from the system operator. Thus, overall productivity should correspond to the expected utility of users' contributions.

In the context of submitting new items to an information repository, we will assume that a user derives utility when an item she submits is added to the repository. In a social voting process, we assume that a user derives zero or negligible utility for submitting an item that does not receive sufficient votes to be added to the repository. (Note that there are arguments for this sort of "rejection" leading to *negative* utility. We acknowledge that a utility loss is possible, but do not consider it in our analysis.)

In a wiki process, when a person contributes by submitting an item, she immediately realizes

a valued outcome – the item that she submitted is immediately added to the repository. However, in a social voting process, others in the community must also take action (by voting for the item) in order for the contribution to produce the desired outcome. It is natural to think of a vote as a selection process – a tool to separate good from bad, or wanted from unwanted – so users should naturally expect that not every item will receive sufficient votes to be added to the repository. Furthermore, typical time preference discount functions suggest that people will strongly favor the instant gratification afforded by a wiki process compared to the delayed gratification in a social voting process [42].

Therefore, due to the inherent delay and uncertainty involved in a social voting process, we believe that users will perceive a lower expected utility of submitting an item to a social voting process than that of submitting to a wiki process. As a result we expect that users will be more motivated and productive when a wiki process is in place.

H1a *Work-Quantity-Submitted*: A wiki process will yield more items submitted by users than a social voting process.

H1b *Work-Quantity-Added*: A wiki process will yield more items added to the repository than a social voting process.

To successfully add an item to a repository in a social voting process, it is necessary to garner support for that item from peers. One way to improve the likelihood that others will vote for the item is to make the submitted item more attractive and high-quality than other submitted options. For instance, a Reddit link with a descriptive (and perhaps slightly embellished) title probably receives more attention from potential voters than one with a bland or uninformative title.

In contrast, in a wiki process, information quality is less important since it is unnecessary to elicit votes from fellow group members. A submitted item is added to the repository regardless of the entry's completeness or quality (though it is possible that extremely low-quality submissions are later removed through other post-hoc processes).

Since the CEM predicts that people will strive to maximize their expected utility gain in the context of their group task, we believe that users will submit items with higher information quality in a social voting process.

H2 *Work-Quality*: A social voting process will produce higher information quality in initial submissions than a wiki process.

It is reasonable to expect that in a social voting process, people will tend to vote for things that they have some level of familiarity with. Therefore, choosing to submit an item that is more "mainstream" should yield a greater likelihood of that item receiving sufficient votes to be added to the repository, and thus, produce greater expected utility to the submitter. On the other hand, somebody who is submitting an item to a wiki process does not need to be as concerned with the group's familiarity with the item. Therefore, we expect that a wiki process will tend to yield items that are more obscure and that have less overall community interest.

H3a *Work-Interest-Submitted*: Items submitted to a social voting process will be of higher community interest than items submitted to a wiki process.

H3b *Work-Interest-Added*: Items added to a repository via a social voting process will be of higher community interest than items added to a repository via a wiki process.

Many social production communities struggle with deviant behavior by some subset of the community. A common type of deviant behavior is submitting undesirable items to a repository, including spam, irrelevant content, or other forms of inappropriate content (e.g., [160, 122]).

The outcome valence aspect of the CEM predicts that such behavior will be more common in a wiki process than a social voting process. In a social voting process, people are unlikely to vote for inappropriate submissions, and thus, those items will have little impact. However, in a wiki process, there is more opportunity for people to be exposed to inappropriate contributions; thus, there is more motivation for deviant behavior.

H4 *Bad-Behavior*: A social voting process will produce less deviant behavior than a wiki process.

Finally, we expect that users will generally be happier with a wiki process, despite the possibility of lower-quality information, items with low community interest, and more visible deviant behavior. That is, we hypothesize that the utility loss from these possible drawbacks and annoyances is outweighed by the utility gain from the wiki process's instant gratification and lack of vote uncertainty.

H5 *User-Satisfaction*: Users will be happier with a wiki process than a social voting process.

4.3 Methods

In this section, we describe the experiment that we conducted on the movie recommender system MovieLens. For the experiment, we modified the process used by MovieLens for adding new movie entries to the database.

The existing method of adding new movies to MovieLens' database was distinctly *not* community driven. Users had the ability to suggest titles to be added, but the decision of what movies to add was ultimately made by the MovieGuru, a volunteer staff member with unilateral control over the movie inclusion policy. Some users found the established policy quite restrictive; for instance, the MovieGuru refused to consider foreign-language movies that are not released widely within the United States.

Note that in [28], Cosley et al. performed a short-term experiment that prompted users to review randomly-selected titles from a pool of movie suggestions, correct or populate information as needed, and decide whether to accept or reject the suggestion based on the established policy. This briefly introduced a limited and indirect form of community curation in MovieLens. While users could decide whether the suggestions they were reviewing would be added to MovieLens, they could not choose which specific suggestions to review, nor could they view the list of outstanding suggestions. As a result, users still did not have the ability to directly add titles of their own choosing.

4.3.1 Experimental Conditions

To make the movie curation process more community-driven, we designed two simple mechanisms to allow MovieLens users to directly add titles to the database. We used a between-subjects design where each MovieLens user who logged in during our experimental period was randomly assigned to one of two experimental conditions. Each condition was associated with one of the two mechanisms. As implied by our hypotheses, the two movie addition mechanisms are based on Wikipedia's wiki process and Reddit's social voting process, and are described in further detail below.

The wiki process, hereafter referred to as WIKI, allows users to add titles to MovieLens without requiring any review from staff members or fellow users. While we did not explicitly mandate or suggest it, we expected that users would view recently-added movies and help check entries for incorrect data or questionable titles. This "patrolling" behavior has been observed in

Figure 4.1: Movie title submission interface used by subjects in both conditions.

social production communities such as Wikipedia ([160, 122]).

The social voting process, VOTE, requires users' submissions to first undergo a community voting process before being added to MovieLens. Users can view others' submissions and vote for ones they believe are worthy of inclusion in MovieLens. The voting mechanism was a simple form of approval voting in which users could vote for as many titles as they wished. Titles that received a vote from at least five users were *promoted* and added to the MovieLens database. To prevent biases from groupthink or bandwagon effects [78], the voting interface did not display vote counts and provided no way to sort the list of submissions by number of votes received.

Figures 4.1 and 4.2 show screen shots of the movie title submission and voting interfaces used by subjects.

To provide users with a larger set of movies that they could add, we relaxed MovieLens' inclusion policy to permit additional classes of movies that were previously disallowed by the MovieGuru. These include made-for-TV movies, and foreign-language movies that did not have theatrical releases in the United States.

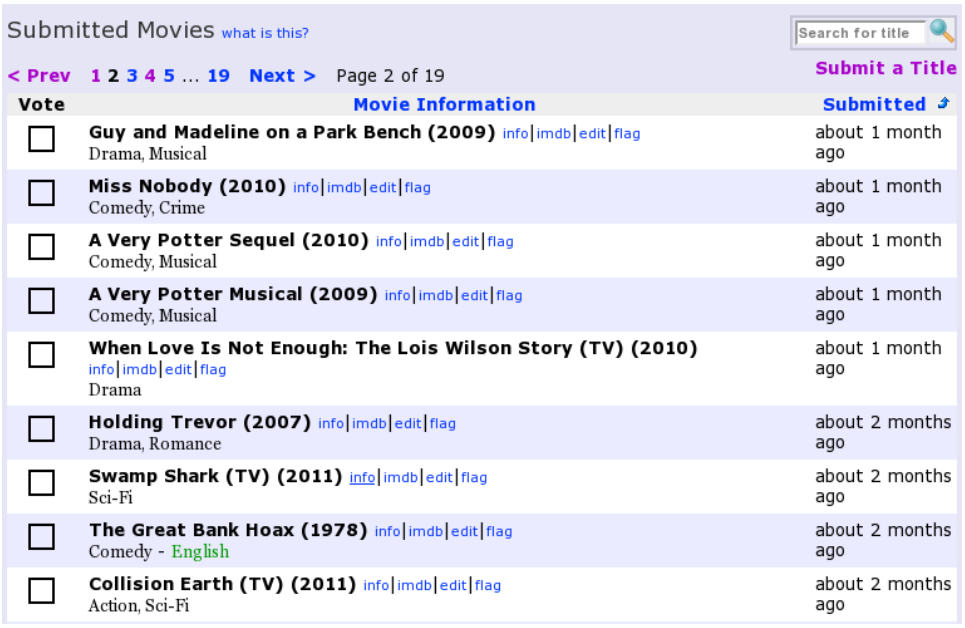

Figure 4.2: Movie title voting interface used by subjects in the VOTE condition.

In both conditions, users were allowed to make corrections to movie submissions and existing database entries as they saw fit. This is consistent with MovieLens' prevailing behavior for existing entries. We also allowed users to "flag" titles that they felt did not belong in MovieLens. Users flagging a title were asked to provide a reason (e.g., "duplicate entry" or "pornography"). A MovieLens staff member evaluated flagged titles and removed ones that did not meet MovieLens' guidelines.[1]

4.3.2 Experimental Design

We struggled to find a fair way to structure the experiment that did not give one condition an inappropriate advantage over the other in which titles were available for submission.

First, we considered a parallel design in which both conditions were simultaneously active. However, we note WIKI has a natural "speed" advantage in that only one person needs to take an action before a title is added to MovieLens, as opposed to multiple people in VOTE (one person must submit the title and vote for it, then several others must also vote for it). It was

[1]During the experiment, the flagging functionality was used 29 times. The vast majority of these were due to concerns about titles that would have been disallowed under the MovieGuru's policy. Only one title (a television series) was removed via the flagging process.

unclear how to address the case where a WIKI user added a movie that was currently under consideration by VOTE users. Thus, we did not believe allowing both the WIKI and VOTE conditions to be active simultaneously was fair, especially to the VOTE condition.

We also looked at the possibility of a "separate worlds" parallel design where each group was segregated into its own "copy" of MovieLens, thus allowing the WIKI and VOTE mechanisms to operate independently of each other. However, existing interactions among users in MovieLens' discussion areas could allow astute users to discover the manipulation (depending on how the manipulation is implemented, users may find that half of their colleagues have disappeared from active discussions, or that some people are talking about movies that do not exist in their experimental condition), leading to possible confounds and awkward user support situations. Furthermore, such a design presented significant technical implementation challenges.

Next, we considered a serial design – allowing one condition to run for several weeks, then the second condition for the same amount of time. This addresses the "race conditions" present in the parallel design, but is still problematic. It gives rise to an ordering effect where the first condition that is active will have more movies available to add than the second condition. For instance, it is possible that the first group will add many popular and well-known movies, thus leaving mostly obscure movies for the second group. Such ordering effects would make the comparison between mechanisms suspect, particularly with regard to our hypotheses about community interest.

We settled on a design that alleviates the issues of both the parallel and serial designs. We used a variation of the serial design that alternated between the WIKI and VOTE conditions on a weekly basis. Also, the number of titles that could be added to MovieLens by each group was arbitrarily capped at fifty titles per week. Users in the VOTE condition could submit more than fifty titles per week for the voting process, but the voting mechanism was altered to allow at most fifty movies to be promoted to the MovieLens database per week.

Per-user submission limits were also put into place to prevent small groups of prolific users from having disproportionately large effects on our results. Each user was limited to two title submissions per day and five submissions per week. Without such limits, we expected that the distribution of movie submissions across users would be extremely skewed toward a few very active "power users." Such distributions have been observed in other SPCs [116, 83, 122], and previous MovieLens experiments have also utilized per-user limits to downplay the effects of power users [28].

We believe this alternating serial design sufficiently avoids the "race condition" issues from the parallel design, and also limits the impact of any ordering effect. As we carried out the experiment, we did find that this design frustrated a small number of users because their ability to contribute to the movie database came and went, seemingly at random from their perspective, but this was a small price to pay to have a fair comparison between the conditions.

We ran the experiment for a total of six weeks. Each of the conditions was active for three weeks each, starting with the VOTE condition. Users who logged in during a week when their assigned experimental condition was "active" were invited to help contribute to and maintain the movie database. After the experimental period, we administered a short survey to all users who clicked on at least one movie submission or movie voting link. The survey asked users about their satisfaction with MovieLens, their reactions to community-driven maintenance of the movie database, and their motivations for contributing or, if applicable, for browsing others' movie title submissions and voting.

4.4 Results and Analysis

We now describe the results of our experiment and whether they support each of our hypotheses. A statistical summary of the titles submitted by users in each condition is shown in table 4.1. The variables in the table will be described as we step through each of the relevant hypotheses in this section.

4.4.1 Submission Quantity

Users in the WIKI condition submitted more movies than users in the VOTE condition, both on a per-user average, and in aggregate. In the WIKI condition, 46 users submitted 134 titles, and in the VOTE condition, 58 users submitted 121 titles. However, the difference in per-user average is only marginally significant (t-Test, $p = 0.092$). Therefore, the results indicate limited support for **H1a *Work-Quantity-Submitted***.

Of the 121 titles submitted by VOTE users, 69 received a sufficient number of votes to be promoted and added to the MovieLens database. Here, the results indicate that **H1b *Work-Quantity-Added*** is supported. WIKI users were, on average, significantly more productive than VOTE users in submitting titles that are (eventually) added to MovieLens (t-Test, $p < 0.001$).

It is interesting to note that despite these results, there were actually fewer WIKI users than

Variable	WIKI	VOTE	V-QUEUE	V-PROMO
Submission Quantity — Section 4.4.1				
Submitters	46	58	–	–
Titles	134	121	52	69
Submitted Information Quality — Section 4.4.2				
PctBad @ Submission	44.8%	51.7%	64.7%	40.6%
PctBad @ Promotion	44.8%	–	–	21.7%
PctBad @ Promotion + 2 Days	8.2%	–	–	10.1%
Movie Properties — Section 4.4.3				
IMDb Popularity - Mean	1,893	1,828	952	2,476
IMDb Popularity - S.D.	2,769	2,557	1,221	3,057
Log IMDb Popularity - Mean	6.93	6.87	6.29	7.31
Log IMDb Popularity - S.D.	1.17	1.16	1.13	0.99
IMDb Likability - Mean	6.97	6.91	6.92	6.90
IMDb Likability - S.D.	1.02	0.92	0.87	0.96

Table 4.1: Properties of movie titles submitted by users during the experiment. **WIKI** and **VOTE** are the titles submitted by users in the two experimental conditions. **V-QUEUE** are titles submitted in the VOTE condition, but that failed to receive sufficient votes for promotion and thus, remained in the voting queue. **V-PROMO** are titles submitted in the VOTE condition that received sufficient votes for promotion.

VOTE users who submitted at least one movie title. We believe this is because of our limits on how many movies could be submitted per day and per week. Some WIKI users clicked on the movie submission link, but were rebuffed because the submission limits had already been reached for that day or week. This suggests that highly-motivated WIKI users were able to monopolize a sizable portion of the daily and weekly limits, thus reducing the opportunity for other users to contribute.

We also found that WIKI users who submitted titles were significantly more active in their use of MovieLens than their VOTE counterparts. An average WIKI submitter had rated 1,118 movies, while an average VOTE submitter had rated 774 movies (t-Test, $p = 0.027$). It appears that the WIKI condition was more effective than the VOTE condition at motivating MovieLens' power users to contribute. However, when asked about their level of motivation for helping maintain the movie database, we found no statistically significant difference between WIKI movie submitters and VOTE movie submitters

4.4.2 Submitted Information Quality

To estimate the quality and completeness of the submitted movie information, we used a metric similar to the $Nfields$ metric used in [28]. We looked for movie information fields that the submitter left blank or formatted incorrectly. For a given set of submitted movie entries, we define a metric *PctBad* as the percent of entries that have at least one blank or incorrectly-formatted information field. This is an imperfect quality metric since it cannot detect some problems such as misspellings or incorrect information. However, simple syntax-based metrics such as this are generalizable to other domains and are easy to implement and work with.

We found that there was no statistically significant difference in the initial quality of movie information submitted in the two conditions. *PctBad* for titles submitted by WIKI users was 44.8%, versus 51.7% for those submitted by VOTE users (Chi-square, $p = 0.3313$). The presence of an explicit peer review and approval mechanism did not appear to affect the initial information quality in users' submissions. Therefore, our findings do not support **H2** ***Work-Quality***.

On the other hand, we did find a significant difference in initial quality between movies in the VOTE condition that eventually received five votes and ones that did not. *PctBad* for VOTE movies that were promoted was 40.6%, compared to 64.7% for movies that were not promoted (Chi-square, $p < 0.02$). Perhaps users were more likely to vote for entries that had complete information. Alternatively, some users may have recognized that certain movies they were submitting for consideration would be unlikely to receive sufficient votes, and thus did not expend the effort to properly populate all the information.

The presence of the social voting process in the VOTE condition allowed for some quality improvements in submitted titles before being promoted. Some users who regularly participated in the voting process also took the time to fix mistakes and fill in omitted information. This allowed titles in the VOTE condition to have higher information quality than titles in the WIKI condition at the time the movie was made available in MovieLens: *PctBad* at promotion-time was 21.7% in VOTE, versus 44.8% in WIKI (Chi-square, $p < 0.01$). However, this disparity closed quickly as MovieLens users saw the newly-added entries and made improvements to them. Entries in both conditions converged to a comparable stable state of information quality within two days of the time they were promoted to MovieLens (Chi-square, $p = 0.61$).

4.4.3 What Got Submitted?

Next, we examine the properties of the movies submitted by each group. To estimate the community interest in submitted movies, we first note that there are multiple ways to measure the "interestingness" of a movie. Two orthogonal dimensions that we will consider are popularity and likability. Popular movies are ones known by many people, while well-liked movies are ones that are generally considered good. Popular movies are not necessarily well-liked (e.g., high-profile box office "bombs" like *Battlefield Earth* or *Waterworld*), and vice versa (e.g., obscure Academy Award nominees).

To operationalize these measures, we use ratings information gathered from the Internet Movie Database (IMDb). We define a movie's number of ratings and average rating as measures of the popularity and the likability of that movie, respectively. Since the distribution of number of ratings is highly right-skewed, we apply a log-transform to make the variable more normally distributed.

Within the VOTE group, we found that users tended to vote for more popular movies. Promoted movies (those that received the requisite five votes to be added to MovieLens) were significantly more popular than ones that did not receive sufficient votes (t-Test, $p < 0.001$). On the other hand, there was no significant difference between promoted and non-promoted movies in movie likability (t-Test, $p = 0.92$); that is, whether a movie was well-received by the public appeared to have no bearing on voting activity.

Interestingly, there was no statistically significant difference in either metric between movies submitted in the WIKI and VOTE conditions. The submitted movies were of comparable popularity (t-Test, $p = 0.71$) and likability (t-Test, $p = 0.75$). The presence of a voting mechanism did not appear to affect users' strategies in choosing what titles to submit for consideration, even though popular movies are more likely to receive votes.

One interpretation of this is that users did not perceive an increase in expected utility by submitting titles that had a higher probability of promotion. In other words, users may have preferred to "scratch their own itch" by submitting an obscure title rather than a mainstream one, even though the latter intuitively has a much higher chance of being promoted in the VOTE condition. As a result, we reject **H3a *Work-Interest-Submitted***.

Next, we compared the conditions in terms of titles that were promoted to the MovieLens database (note that movies submitted by the WIKI group are trivially promoted). We found that movie likability was comparable between the groups (t-Test, $p = 0.73$); however, the VOTE

group's promoted movies were significantly more popular (t-Test, $p = 0.017$). Therefore, the social voting mechanism did lead to more "mainstream" and high-interest items being added to the database, which supports **H3b *Work-Interest-Added***.

4.4.4 Deviant Behavior

Here, we examine the presence and frequency of users behaving in deviant ways that are contrary to community norms. We expected to see some amount of non-constructive behavior in MovieLens, particularly with the WIKI condition where users could freely add whatever titles they wanted directly to MovieLens.

However, we found that users in both conditions were well-behaved. While there were several title submissions that some users questioned because they would have been questionable under the MovieGuru's editorial regime, there were no signs that users were deliberately acting in bad faith. Specifically, we looked for cases where users were providing wrong or nonsensical movie information, submitting obviously inappropriate titles (e.g., pornography or short viral videos), or creating multiple accounts to vote multiple times.

Notably, this lack of deviant behavior is a departure from the results of a previous MovieLens experiment by Cosley et al. in [28] where a moderate amount of "hijacking" behavior was observed. In their experiment, some users did not complete the curation task that they were given. Instead, they "gamed the system" and tried to introduce titles that they personally wanted to be added to MovieLens, even though doing so resulted in movie entries with incorrect metadata.

We believe that this result is due to inherent properties of the WIKI and VOTE mechanisms. Compared to the mechanism offered in [28], the ones we studied are more closely aligned with users' personal interests. Rather than mandating that users make a decision about randomly-selected submissions that they may have never heard of before, these mechanisms provide users more choices in what tasks to do as well as a more direct path to add titles of their own choosing. Therefore, perhaps users in both conditions perceived relatively small utility gains from deviant behavior. While the mechanisms we studied are susceptible to undesirable actions (e.g., vandalism, Sybil attacks, and vote collusion), we did not observe these problems during our experiment, and thus, we found no support for **H4 *Bad-Behavior***.

Name	Statement
MOTIVATION	I felt motivated to help maintain the movie database.
IMPACT	I think my own submissions had a positive impact on the movie database.
GOODJOB	I think that MovieLens users have done a good job so far in adding movies.
ENJOYMENT	I liked being able to help decide what titles are added to MovieLens.

Table 4.2: User satisfaction statements from post-experiment survey. Participants were asked to rate each of these statements on a 5-point Likert scale.

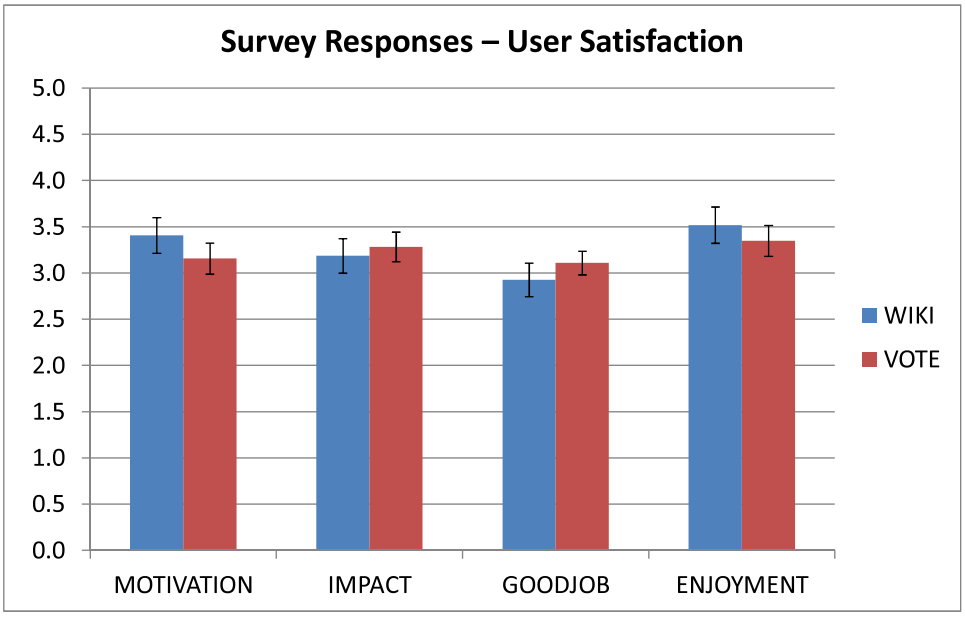

Figure 4.3: Survey responses to user satisfaction statements (see table 4.2 for text of statements). Error bars represent two standard errors in each direction. $N = 27$ for WIKI, and $N = 46$ for VOTE. Differences between WIKI and VOTE are *not* statistically significant.

4.4.5 User Satisfaction

Our post-experiment survey asked participants to rate several statements on a 5-point Likert scale in order to gauge participants' motivation level, perceived individual performance, perceived group performance, and overall satisfaction in community maintenance and curation of the MovieLens' database. The text of the statements is shown in table 4.2.

Users' responses to these questions are summarized in figure 4.3, and indicate that despite our expectations, there was no difference between active participants (users who submitted a

title or voted for a title) in our two conditions for *any* of these four statements. [2] Thus, **H5 *User-Satisfaction*** is not supported.

It is unclear why we observed no significant differences in these self-reported survey results. Perhaps the differences in user satisfaction between the two conditions are simply too small to yield a measurable result, despite the mechanisms – a social voting process, and a wiki process – being diametrically opposed on several dimensions of mechanism design.

Another possibility is that individual differences in attitudes about curation policy clouded our survey data, thus making between-group differences difficult to measure. Responses to open-ended survey questions reveal a wide spread of user attitudes and preferences regarding the MovieLens database and what types of movies it should contain. Some users were positive about the emergent changes to MovieLens' database:

> "My only complaint with this site had always been the amount of movies missing from the database. For the first time in years these movies are being added, and at a fairly quick pace. I also like that the criteria has changed [...], allowing for more foreign films and great TV movies to be added."

> "I don't think anybody should decide whether or not a movie is appropriate for MovieLens. [...] For all I care; bring on Youtube movies and ratings into the database."

However, other users preferred the careful and selective curation that the MovieGuru provided, and were unhappy with the influx of user-submitted titles:

> "Too many movies got into the system that should be removed. Movies still on the festival circuit or foreign films that aren't readily available to the U.S. audience (or other audiences for that matter) should not be in the system."

> "Lots of obscure and irrelevant films from no-names are being added, including TV movies and direct-to-video releases. They dilute the database."

These diverging viewpoints are reminiscent of Wikipedia's inclusionist-deletionist dynamic that we described and studied in chapter 3. The disagreement among members of the MovieLens community regarding which films to include in the database could be a confound that

[2] We obtained qualitatively similar results if we omit users who only voted for titles in the VOTE condition.

obscures our ability to determine which curation mechanism users were happier with. That is, participants' satisfaction with the curation interface and workflow may be colored by their opinions about the other titles that were added to the database, perhaps leading to a general feeling of ambivalence.

4.4.6 New Releases

A surprising result that we noticed during data analysis was that users tended to not submit newly-released films regardless of which condition they were in. During the six week experiment, there were 33 movies that were new "wide releases" in the United States (defined as appearing in at least 600 theaters[3]). Of these movies, users submitted only 11 to MovieLens. There were also 20 new limited-release movies (appearing at fewer than 600 theaters) during our experiment, of which just three were submitted. Furthermore, in many of the cases where new releases were submitted, the submissions occurred well after the theatrical release date: half of new releases were not submitted until at least a week had passed since the movie's theatrical release date.

This phenomenon is potentially damaging to MovieLens as a service because it reduces the service's usefulness to some users. Since MovieLens is primarily a movie recommender system, many of its users use it solely to find movies to watch. If MovieLens does not have new theatrical releases available in its database, then it cannot recommend those movies to users. Therefore, those who use MovieLens to get recommendations for an outing to the theater may find the system to be lacking if newer releases are not added on a timely basis.

These new releases typically have large marketing campaigns (especially "wide releases") and are thus known by a substantial proportion of the movie-watching public. As a result, the intuition might be that the community will be quick to submit these "mainstream" titles, since they represent low-hanging fruit. Users reported similar intuitions about what would happen with new releases. When prompted on the post-experiment survey for comments about what they did during the experiment, two users wrote:

> "New releases are always added, [but] classics and small but notable releases – especially foreign films – sometimes get overlooked."

[3]This definition appears to be a de-facto movie industry standard.

> "...[I work on] the new releases not so much because I watch mainly old movies and others like to race to get the information up anyway."

Upon further reflection and analysis, we believe that the CEM offers a plausible explanation for this result. One of the CEM's predictions is that people will be less motivated to contribute when they believe their work will not be unique when compared to the groups' work. Since there may have been a common intuition that many peers know about new releases, submitting such titles represents a non-unique contribution and is, therefore, less likely to occur.

We found that this was indeed the case – users tended to submit older and more obscure movies that they may have perceived as a unique contribution, rather than new releases that many of their peers likely already know about. When we asked participants about their likelihood of performing various curation tasks, their survey responses indicated that they were significantly less likely to submit new release movies than they were to perform other tasks.[4]

This result suggests we may need to modify the mechanism design to incentivize and encourage people to perform important tasks that they have low motivation for. For example, in MovieLens, we could offer increases in community visibility, submission limits, or vote influence to users who can supply a steady stream of new release submissions.

4.5 Discussion

Our findings lead us to several recommendations and guidelines regarding the design and selection of new-item curation mechanisms in social production communities.

Properties of Added Items. A key difference that emerged between our two experimental conditions was in the popularity of the items that they added to the repository. The social voting process yielded movies that were more mainstream and less obscure than the movies added via the wiki process. Whether this is considered a feature or a bug is a value judgment on the part of the SPC operators and the community. As seen in our explorations of Wikipedia in chapter 3 and in the participant comments shown in section 4.4.5, there can be substantial disagreement regarding this issue. Nonetheless, system designers should carefully consider what shape they want the SPC information repository to have when deciding which curation mechanism to use.

[4] On a five-point Likert scale, respondents indicated they were significantly less likely to *Submit movie information about new releases* than they were to *Fix errors in movie information, such as typos, missing fields, or other mistakes* (t-Test, $p = 0.043$). Respondents were also significantly less likely to *Submit movie information about new releases* than *Submit movie information about titles that I think should be added to MovieLens* (t-Test, $p < 0.001$).

Submission Quantity. Compared to a wiki process, a social voting process resulted in reduced user motivation and productivity, partially due to the lack of instant gratification. Furthermore, a social voting process produces more "wasted" effort as not all submissions will muster a sufficient number of votes from the community. Thus, a wiki process is preferable to a social voting process if one wishes to maximize the net amount of work performed.

Submission Quality. The choice of curation mechanism had a limited effect on submitted information quality. We found that the social voting process offered some opportunity for quality improvements before items were promoted to the repository. However, neither submission-time information quality nor steady state information quality differed significantly between the two conditions. This result is in line with the projections made by Cosley et al.'s models in [28] – that is, the overall value of an SPC's information repository converges to the same level regardless of whether community review occurs before a contribution is added to the repository (albeit at different rates). Therefore, when considering information quality, the primary factor that affects curation mechanism design is the SPC's tolerance to low quality information temporarily appearing in the repository.

Mechanism Design. Curation tasks that are entrusted to users should be designed and framed in a way that aligns users' interests with system and community interests. Perfect alignment may not always be possible; however, designers should endeavor to structure tasks and incentives in a way that encourages desired behavior. Doing so has two benefits.

First, it can help reduce deviant behavior by reducing the incentive to behave badly. For instance, the mechanisms we studied in this work effectively eliminated the "hijacking" behavior seen in [28] by offering users more freedom and more choices instead of mandating completion of specific curation tasks without any reward.

Second, it can help improve the SPC be more efficient in reaching its goals. Disparities that exist between individual contributor interests and system or community interests may lead to information repositories that have diminished usefulness. For example, we found that participants were reluctant to perform an important curation task: keeping MovieLens up-to-date with new releases.

Summary. In this chapter, we have empirically studied and compared two very different collaborative curation mechanisms. The results suggest that the mechanism has a substantial effect on information quantity as well as what types of items are added to the repository. On the other hand, we observed a limited effect on information quality, and no apparent effect on user

satisfaction.

This work has only scratched the surface in exploring collaborative curation mechanism design and its consequences. There exist numerous mechanisms beyond the two that we studied, and we believe that as SPCs continue to grow and mature, more work will be needed to understand how the underlying curation mechanisms will influence their evolution, and how designers and operators can make appropriate choices to help guide the growth of the SPC.

Research Theme 2: Challenges in Collaborative Curation

Chapter 5

Effects of Group Composition on Wikipedia Curation Decisions

5.1 Introduction

We begin our exploration of challenges in collaborative curation by looking at how small groups make curation decisions and resolve conflict in an SPC. As community members collaborate disagreements will inevitably arise, and the participants must make decisions about how to resolve conflict and move forward. Effective decision-making and conflict resolution processes are essential to a healthy community. Flawed processes may lead to poor decisions, which are costly to address—not only can bad decisions increase coordination costs and process losses, they can also alienate users and cause them to leave the community.

5.1.1 Contributions

This chapter describes our explorations of a group decision-making process in the context of Wikipedia. We analyze over 100,000 content curation decisions made by small working groups in Wikipedia, and study how four group composition factors affect the quality of the decisions that are made. Our results lead us to a number of recommendations and implications for the design of SPCs.

5.1.2 Related Work and Research Questions

Group decision making is a rich area of research that has been studied extensively in multiple disciplines, including social psychology, economics, and political science [56, 94]. One limitation of the existing literature is that much of it focuses on group composition factors that affect performance in physical face-to-face settings. A goal of the current work is to learn how these factors apply in computer-mediated communication (CMC) settings where group members are in an environment that lacks nonverbal and paraverbal cues, often working asynchronously and with anonymous or pseudonymous peers. Because of these differences, we cannot assume that the findings from offline groups will apply in an online context.

There have been been numerous comparisons of group performance between groups that use CMC and those that use face-to-face communication (e.g., [82, 142]), but many are limited to studying the broad performance differences between groups in offline and online environments. A meta-analysis by Baltes et al. examined 27 such studies and found that CMC groups generally underperformed face-to-face groups [8]. CMC groups took longer to make decisions, made worse decisions, and had lower member satisfaction. The meta-analysis found several factors that influenced the effectiveness of CMC groups, including anonymity, group size, and task type. We seek to expand on this knowledge and find ways for designers of SPCs to improve their decision-making processes.

We now review the literature on group decision making and conflict resolution, focusing on and highlighting several group composition factors that influence decision-making acuity in face-to-face settings. These factors will become the basis for our research questions in the current work.

Social psychologists have found that group size affects the dynamics of conflict resolution processes in small groups. People in larger groups are prone to escalating conflict and are less likely to cooperate with one another [94]. Larger groups also suffer from process losses, which may reduce efficiency and performance [144]. On the other hand, in *The Wisdom of Crowds*, Surowiecki suggests that in many cases, large and diverse groups can make better decisions than individuals or experts. Small groups risk making worse decisions because they may lack relevant information or a diverse range of viewpoints [149].

In an online context, large groups may be more manageable thanks to asynchronous communication tools that allow participation without requiring that everybody be simultaneously present and paying attention. Intelligent user interfaces allow long discussions to be easily

browsed or searched. However, the often impersonal nature of online interactions may serve to encourage uninhibited and antisocial behavior [82, 142, 147], which may further reduce cooperation and increase conflict in large groups. This brings us to our first research question:

RQ1 ***Group Size.*** How does group size affect decision quality in online communities?

A crucial part of any decision-making process is defining the group that is responsible for making the decision. Ideally, a decision-making group should be a representative subset of the organization or community that needs the decision to be made. Traditionally, groups have been created in a top-down manner by an authoritative figure (i.e., a manager, or in academic studies, the experimenter). However, some groups, such as working groups or ad hoc committees, can be self-forming.

The manner in which self-formed groups attract and recruit participants can have a profound effect on group composition, which can in turn influence decision quality. For instance, group members may naturally choose to solicit those in their own social networks. Because social networks exhibit homophily and tend to be a source of behavioral homogeneity [103], the resulting group composition may be skewed toward particular attitudes or preferences that are not representative of the community as a whole. Self-formed groups are increasingly common, especially in online communities [56], and in the current work, we look at one aspect of biased group recruitment in an SPC, and explore how it affects decision quality.

RQ2 ***Group Formation.*** How does biased group formation affect decision quality in online communities?

The members of a decision-making group are likely to have differing levels of experience working with the organization or community. Some participants may be oldtimers with substantial experience, while others may be newcomers who are still learning about their roles. The group diversity literature suggests that such diversity can be both good and bad. The informational perspective hypothesizes that heterogeneous groups do better because they have a broader range of knowledge, skills, and opinions to draw from. Newcomers provide new ideas and perspectives, while oldtimers provide experience and structure. On the other hand, the social categorization perspective suggests that diversity is harmful because people use the differences to categorize group members into subgroups, which can lead to increased conflict and an adversarial “us versus them” dynamic [156].

In online communities, the effects of tenure diversity may be confounded by the fact that indicators of tenure and status are not always made salient. Many have hypothesized that masking these indicators reduces the effect of social categorization and status inequalities, which, in turn, can help equalize participation levels, promote communication openness, and improve decision quality. However, study results have been equivocal [8, 22]. Our next research question seeks to explore further the role of newcomer participation and tenure diversity on decision quality.

RQ3 ***Experience.*** How does newcomer participation and tenure diversity affect decision quality in online communities?

Finally, decision-making groups typically have an administrator or leader who is responsible for identifying and executing the group's decision. In some cases, that person may be partial to a particular outcome, and may use their influence to steer the group toward that outcome. Decisions made under such conditions are suspect because valid arguments and viewpoints may be ignored.

For instance, in [149], Surowiecki describes a crucial decision made by the NASA Mission Management Team during the space shuttle Columbia's final mission. The team met to decide whether to more thoroughly investigate the possibility that the shuttle sustained severe damage during launch. Surowiecki presents evidence that the team's leader had already made up her mind (before the meeting) that the damage was inconsequential, and deflected and downplayed issues brought up by engineers during the meeting. In essence, the leader ignored a number of valid concerns, leading to a flawed decision-making process and a decision that arguably resulted in the loss of the shuttle and its crew. Similar instances of leader bias have been reported in judicial decisions [120].

In SPCs, biased administration may be even more of an issue as self-formed groups become more prevalent. Administrative roles in such groups are often volunteer-based or even self-appointed, which may be susceptible to selection biases, since volunteers may choose to seek power in areas where they have strong pre-formed opinions. Also, even if administrators endeavor to perform their duties in an impartial manner, they may still be affected by subconscious or hidden biases [55]. In the current work, we study the effects of apparent administrative bias on decision quality.

RQ4 ***Administrative Bias.*** How does biased group administration affect decision quality in online communities?

5.1.3 Decision Making in Wikipedia

Our overarching goal in asking these research questions is to understand how decision-making processes in SPCs work and to learn how to improve their effectiveness through better processes, software tools, and intelligent interfaces. In the present work, we explore these questions in the context of one of the largest SPCs in the world: the English Wikipedia. With millions of contributors and articles, countless decisions must be made every day to keep the encyclopedia running smoothly.

There has been substantial research in how Wikipedians successfully manage such a large community. Forte and Bruckman examined Wikipedia's self-governance, noting that there has been an increasing level of decentralization in its decision-making processes; decisions that were once reserved for founder Jimmy Wales have become entrusted to the community [40]. Other research has examined more specific aspects of Wikipedia's decision-making processes, including how specialized tools enable vandal fighters to make decisions more efficiently [45], how Wikipedia's user promotion decisions compare to stated policy [19], and how the community decides which articles to feature on Wikipedia's front page [158]. While each of these decisions features qualitatively different workflows and processes, they are all distributed community-driven ways of arriving at a consensus that strive to avoid giving excessive authority to any single entity. In the present work, instead of focusing on decision workflows, we look at group composition and its effect on decision quality.

One of the most important content curation decisions that an SPC must make is to define the scope and breadth of the community's efforts. In chapter 3, we saw that as many as one-third of new articles are deleted, often because Wikipedians believe that the articles are about topic that are not sufficiently notable or important for the encyclopedia. In this work, we address our research questions by analyzing the processes that small groups of Wikipedians use to decide whether to delete an article, and by looking at how different group composition factors influence the quality of their decisions.

5.2 Data and Methods

5.2.1 Article Deletion on Wikipedia

In Wikipedia, deleting an article involves an extensive set of processes that is uncharacteristically *not* wiki-like. Recall from our discussion in section 3.6 that while anyone can create and edit Wikipedia articles, most users cannot delete articles. Ordinary editors are limited to proposing and discussing deletions via one of several processes.

In this chapter, we are interested specifically in Wikipedia's *Articles for Deletion* (hereafter referred to as *AfD*) process because it involves the community coming together and making a collective decision about what to do with a specific article that somebody has nominated for deletion. We choose to study decisions made by this particular process because they are organized in a relatively standardized format that is amenable to automated coding, they occur frequently enough for our quantitative methods to be effective, and they involve self-formed groups with a sufficiently wide variety of compositions to address our research questions.

The AfD process is used when there is possible disagreement over whether an article belongs in Wikipedia. To begin an AfD discussion, a user nominates an article to be deleted and provides his or her reasoning. Then, interested members of the community spend five or more days discussing the deletion. Finally, a neutral administrator examines the group discussion and determines what the community has decided to do. The administrator then takes the appropriate action, and closes the discussion. The typical outcomes are to delete or to keep the article.

A typical AfD discussion and decision is shown in figure 5.2.1. Here, the user Merope has nominated the article *Lighthouses in Spain* for deletion because he believes that it is a trivial list of information that adds little value to Wikipedia. Over the next eight days, several others discuss whether the article should be deleted: Kwsn, C.Logan, and JForget favor deletion, while Dhaluza, Dhartung, Steve Hart, and Sjakkalle are opposed. Finally, administrator Akhilleus determines that the community has reached consensus to keep the article. He announces the result, thereby closing the discussion.

To the casual observer, it may appear as though the closing administrator is merely tallying up how many participants "voted" for each outcome because the participants structure their arguments in a vote-like format. Each argument is prefixed with a clear and brief summary that is visually distinct and easily counted (e.g., "**Keep**" or "**Delete**"). However, Wikipedia's

The following discussion is an archived debate of the proposed deletion of the article below. ***Please do not modify it.*** *Subsequent comments should be made on the appropriate discussion page (such as the article's talk page or in a deletion review). No further edits should be made to this page.*

The result was **keep.** --Akhilleus (talk) 20:43, 28 June 2007 (UTC)

Lighthouses in Spain [edit]

Lighthouses in Spain (edit | talk | history | links | watch | logs) – (View log)

zomg listcruft!!!111! Erm. Sorry. Wikipedia is not an indiscriminate list of information. -- Merope 17:42, 20 June 2007 (UTC)

- **Delete** Category is there already, no need for a list. Kwsn (Nil) 17:51, 20 June 2007 (UTC)
- **Delete** - Per Kwsn. Why do we need two pages to do the work of one?--C.Logan 18:00, 20 June 2007 (UTC)
- **Delete** since the category exist, lots of red links too. But there at list 15-20 other list similar to that, so I guess most of them are listcruft as well.--JForget 19:34, 20 June 2007 (UTC)
- **Keep** This is Cruftcruft. First, the category only lists articles created--many of the lighthouses on the list will never have stand-alone articles, and we usually merge these to a list. Also lighthouses are prominent geographic landmarks, important in both marine navigation, and human culture. Each one must be unique for identification, and the unique characteristics are often associated with the adjacent settlements. Dhaluza 00:56, 21 June 2007 (UTC)
- **Keep**, a perfect example of a list that does what a category cannot, show articles that are not yet created. --Dhartung | Talk 04:38, 21 June 2007 (UTC)
  - **Comment** I'll give you that one, but how can the lighthouses be verified? That's a big concern of mine. Kwsn (Nil) 17:21, 22 June 2007 (UTC)

    Look on a map. We don't delete articles because you can't verify it without getting out of your chair. Dhaluza 12:01, 28 June 2007 (UTC)
- **Note**: This debate has been included in the list of Spain-related deletions. -- John Vandenberg 09:16, 21 June 2007 (UTC)
- **Weak keep**, since we're using lists on WP, this one serves its purpose. -- Steve Hart 14:55, 25 June 2007 (UTC)
- **Keep**. Lighthouses are important features in ocean navigation, and therefore important geographical landmarks, often receiving a pretty prominent mark on maps. Sjakkalle (Check!) 10:48, 28 June 2007 (UTC)

The above discussion is preserved as an archive of the debate. ***Please do not modify it.*** *Subsequent comments should be made on the appropriate discussion page (such as the article's talk page or in a deletion review). No further edits should be made to this page.*

Figure 5.1: A typical Wikipedia Articles for Deletion (AfD) discussion (from http://en.wikipedia.org/w/index.php?oldid=141246516).

guidelines state that AfD, along with most of its other decision-making processes, are *not* vote-based. Instead, Wikipedians expect administrators to carefully study the arguments and determine whether the participants have reached a "rough consensus" in deciding what to do.[1]

Wikipedians refer to these vote-like statements as "!votes" (read as "not-votes") as a tongue-in-cheek reminder that while the discussions may resemble votes, voting is not actually taking place, and that opinions that are not accompanied by valid reasoning may be disregarded. For brevity, we will adopt similar nomenclature, referring to discussion participants as ***!voters*** and their preferred outcomes as ***!votes***. Also, when the meaning is clear from context, we use the term "AfD" to refer to specific instantiations of the Articles for Deletion process, rather than the process itself.

5.2.2 Measuring Decision Quality

A key part of exploring our research questions is knowing whether the AfD decisions being made are good. The usual scientific approach here might be to identify what factors Wikipedians use to evaluate whether an article belongs in Wikipedia, and to operationalize them as metrics that estimate these factors. For instance, in chapter 3, we found that the plurality of article deletions occur because the articles are about topics that do not sufficiently meet Wikipedia's notability guidelines. We defined metrics to estimate the notability of an article topic to help us look for evidence of whether deletions were generally consistent with notability guidelines.

In principle, we could use these same metrics to help classify decisions as good or bad by looking for keep decisions made on low-notability articles, or delete decisions made on high-notability articles. However, we believe that such an approach is awkward for two reasons.

First, metrics of this nature are imprecise. The notability metrics we used are noisy and are difficult to apply definitively in individual cases. Wikipedians are aware of these metrics and do not consider arguments based solely on them to be valid. We believe that most extrinsic metrics of this type will be similarly unsuitable for rendering judgment on individual decision correctness.

Second, even if we had precise and accurate metrics, we are still left with the problem of defining a value as the threshold that an article must meet to avoid deletion. This is problematic because there is no gold standard. Part of Wikipedia's ethos is to allow its community to make its own decisions about content, style, and governance. There is no wrong decision as long as

[1] http://en.wikipedia.org/wiki/WP:NOTAVOTE

it was made in good faith as a way to move forward with the overarching goal of producing a free, high quality encyclopedia. It is difficult for us as outsiders to justify declaring that some AfD decisions were "wrong" just because a metric that we invented said so.

Instead, we measure decision quality by observing feedback in the system itself and looking for evidence that the community believed that a decision it previously made was incorrect. In the context of Wikipedia AfDs, we look for decisions that are reversed; that is, we find cases where an article is:

- deleted via AfD, but is re-created at a later date, or
- kept via AfD, but is deleted at a later date.

These reversals can occur through a variety of mechanisms. For instance, the decision may have been reversed due to a formal appeal lodged at one of Wikipedia's dispute resolution channels, or an informal conversation among the involved participants. Alternately, a bold user or administrator might have simply taken the initiative to reverse a decision that he or she felt was incorrect. Since decision reversals can themselves be reversed upon community scrutiny, we only consider reversals that "stick"; that is, reversals that are persistent and are not undone. To help avoid cases where decision reversals may be due to policy changes or other long-term changes in the ecosystem, we only consider cases where an AfD decision is reversed within one year as being an indicator of a flawed decision.

In addition, we do not consider cases where articles are re-created as a redirect (a pointer to another article) to be a flawed deletion decision. Such cases may occur if Wikipedia has an article that is topically related to the deleted article. As a courtesy to readers, editors may opt to re-create the article as a redirect to the related article, especially if the deleted article's title might be a common search term.

We acknowledge that this is an imperfect method to measure decision quality. Not all bad decisions will be fixed by the community, and not all reversals are the result of flawed decisions. However, we feel that this approach represents an effective microscope into which decisions are of questionable quality, and allows us to study them without requiring us to impose our own judgment about the community's decisions.

5.2.3 Data Sources

To collect the requisite information to explore our research questions, we used the Wikimedia Foundation's data dumps[2] and the Wikimedia Toolserver[3].

Current Versions Dump. We used the current versions dump, which contains the current text of every Wikipedia page, to obtain the text of all archived AfDs. Using the *mwlib* library to parse the wiki markup, we wrote a program that extracted the key information from each AfD: the article being discussed, the nominator's name, the participants' names and !votes, the closing administrator's name, and the decision. We found that about 1% of AfDs did not appear to be in the de-facto standard format shown in figure 5.2.1; these ill-formed AfDs are excluded from our analysis.[4]

To check the program's correctness, two people independently examined 48 random AfDs and noted any errors in the extracted data. The judges found errors in 3.8% of the pieces of collected information. Many errors involved the program misidentifying a participant's !vote, and were due to people expressing their !vote in unconventional ways, or making complicated and nuanced arguments that could not be classified easily. We felt that an accuracy rate of over 95% was acceptable for a simple parser, and did not believe that more complex techniques such as sentiment analysis would be worth the added cost.

Metadata Dump and Event Log. The remainder of our data came from two sources: the historical revision metadata dump and the event log dump. The revision metadata dump tells us when each Wikipedia edit occurred, and who made each edit. The event log tells us when pages were deleted, restored, or renamed. We used these data sources for three purposes.

First, the data helped us verify and refine our automated coding program's outputs. Each AfD event (nomination, !vote, or closure) should correspond to an edit to the AfD discussion page at the time indicated in the user's signature. Using the revision metadata, we checked whether this was true for our collected data. If not, we checked whether a simple username or time correction would make the event consistent with the metadata (some Wikipedians' signatures contain alternate versions of their username or a non-UTC time). If that failed, we omitted the suspect event from our analysis. Additionally, we used the data to verify whether our program correctly assessed each AfD's result. For instance, if an AfD resulted in a delete decision,

[2] http://en.wikipedia.org/wiki/WP:DUMP

[3] http://meta.wikimedia.org/wiki/TS

[4] Many of these are corner cases such as unfinished AfD nominations, cases where the AfD discussion itself was deleted or redacted, or discussions that were arranged in some unrecognizable way.

then a corresponding deletion should appear in the event log immediately following the AfD's closure.

Next, the data allowed us to detect whether an AfD decision was reversed, which, as described in section 5.2.2, is our indicator of an incorrect decision. For example, if an AfD resulted in a keep decision, we looked for evidence of a reversal by searching the event log for a deletion occurring after the discussion was closed. If we found one, we would then also search the data for the article's subsequent re-creation to determine whether the reversal was sticky.

Finally, we used the metadata to compute other metrics about each AfD that we use in our model of decision quality. Since the metadata dump omits deleted articles, we used the Wikimedia Toolserver as needed to obtain metadata about deleted articles. We will describe these metrics in sections 5.2.5 and 5.3.

5.2.4 AfD Data Set

Our data set contains 158,733 AfDs occurring between January 1, 2005 and April 1, 2009. We chose to exclude discussions starting before January 1, 2005 because many of them were not in the format depicted in figure 5.2.1, and thus could not be processed by our automated coding tool.

Most AfD decision-making groups are small—the median number of !voters (not including the nominator) is four. Because we are interested in studying *group* decision making, we omit 12,997 AfDs that had zero or one !vote from our analysis.

A majority of AfDs, 68%, result in a decision to delete the article, while 25% result in the article being kept. The remaining 7% represents a variety of uncommon outcomes, including merging the article's content into another article, redirecting readers to another article, or moving the content to another wiki. Because it is unclear how to identify whether such decisions are reversed, we discard this 7%, which leaves a total of 135,461 AfDs in our analysis. Of these, we found that 4.67% of delete decisions and 3.52% of keep decisions are reversed.

At first glance, the fact that over two-thirds of discussions result in deletion may make AfD look like an unfairly biased process. However, the disparity is perhaps expected: it is reasonable to believe that someone would only nominate an article for deletion if there were good reasons for doing so (thus, making deletion a likely outcome). Someone who nominates articles haphazardly might face consequences for being disruptive.

5.2.5 Modeling Decision Quality

Our analysis of decision quality uses a logistic regression model. The binary dependent variable is whether the AfD decision is reversed, and the independent variables represent properties of each decision-making group. To account for factors that may correlate with reversals but that are not related to the factors that we are studying, we control for several constructs as described below:

Temporal effects. As Wikipedia and its users age, there may be natural changes in how often decisions are reversed resulting from factors such as policy changes or broad shifts in community behavior. We control for this by including a *DiscussionDate* variable, the date that the AfD discussion was started, in the model, expressed as the number of years after January 1, 2005. Additionally, we control for any effects due to the freshness or staleness of the nominated article, given as *ArticleAge*, the article's age in days at the time it was nominated for deletion.

Strength of consensus. There is reason to believe that decisions made through weak consensus are more likely to be reversed than those with strong or unanimous consensus. To model the strength of consensus, we use the percentage of participants who !voted for the eventual decision (*ConsensusStrength*). A value of one indicates a "unanimous" consensus, while lower values indicate weaker consensus and (perhaps) increased controversy.

Stakeholder impact. The stature and size of the groups affected by a decision and their participation in making the decision may have an impact on whether the decision is reversed. In the context of AfDs, the user(s) who authored the nominated article likely feel the most affected by the debate since the community is considering throwing their work away. We use three variables to control for this: the number of users who contributed to the article before its nomination (*NumEditors*), the experience of the user who created the article (*CreatorEdits*, defined as how many Wikipedia edits the user had made before creating the article), and whether he or she was involved in the AfD discussion (*CreatorVoted*).

Decision outcome. Different decisions may be more or less likely to be reversed depending on factors such as group attitudes, norms, and procedures. In the context of AfDs, reversing a keep decision may require the community to go through the deletion processes again, while reversing a delete decision may entail rewriting the article. These differ in difficulty, formality, and level of community scrutiny. Since these are fundamentally different processes, we believe that our research questions may have answers that depend on which decision was made. To this end, we present our results as two separate models: one for AfDs that resulted in a delete

decision, and one for AfDs that resulted in a keep decision.[5]

5.3 Results and Analysis

Now, using the AfD data set and our models of decision quality, we explore each of our four research questions. In each subsection, we will present analysis of policy or exploratory data to motivate interesting hypotheses, state our hypotheses, explain how we tested each hypothesis, and describe the results from our full models.

Table 5.1 shows our two models of decision quality, along with descriptive statistics about the input variables. Control variables are described in section 5.2.5, and independent variables are described in the following subsections. All variables have a variance inflation factor (VIF) of below 2.5, which suggests that inflated standard errors due to multicollinearity is not an issue [72]. Since the distribution of several variables is right-skewed, often with standard deviations larger than the mean, we apply base-2 logarithms to transform them to approximately-normal variables. These variables are labelled with "log2" in table 5.1.

5.3.1 RQ1: Group Size

We begin by exploring the question of whether, as suggested by research on offline group dynamics, group size affects the AfD decision process. Figure 5.2 indicates that different-sized groups tend to yield different decisions. Large groups make fewer decisions to delete an article than average, while small groups make more decisions to delete an article than average. Figure 5.3 shows that as group size grows, so does the average percentage of people who dissented and !voted against the eventual decision. These relationships suggest that group size has a fundamental effect on how decisions are made, and lead us to believe that it may have an effect on decision quality. We believe larger groups will benefit from additional viewpoints and information, but with diminishing returns since conflict and dissent will also become increasingly prevalent.

H1 *Bigger-Better*: Larger groups will make better decisions than small groups, but with diminishing returns.

[5] We obtained qualitatively similar results performing the analysis using a single unified model with the decision outcome included as (a) a binary variable, and (b) interaction terms with all other described variables. For clarity, we choose to present the results as two separate models.

Model 1: Delete decisions					Model 2: Keep decisions			
Mean	**SD**	β	**OR**	**Variable**	**Mean**	**SD**	β	**OR**
–	–	-1.8233	0.161 ***	Intercept	–	–	-0.6152	0.541 ***
2.08	1.09	-0.0303	0.970 +	DiscussionDate	2.32	1.11	-0.4649	0.628 ***
156	285	0.0035	1.004	ArticleAge (log2)	346	439	-0.0447	0.956 ***
576	3861	0.0352	1.036 ***	CreatorEdits (log2)	2094	8163	-0.0095	0.990
10.67	40.34	0.1409	1.151 ***	NumEditors (log2)	35.6	142	0.2401	1.271 ***
0.0556	0.229	-0.2065	0.813 **	CreatorVoted (0/1)	0.159	0.366	0.4086	1.505 ***
0.909	0.161	-1.9025	0.149 ***	ConsensusStrength (%)	0.801	0.199	-3.4552	0.032 ***
RQ1 ***Group Size*** — Section 5.3.1								
5.43	4.20	-0.1159	0.891 ***	H1: GroupSize (log2)	8.22	7.40	-0.1597	0.852 ***
47.1	187	0.0227	1.023 *	H1: GroupSizeSq	122	725	0.0365	1.037 *
RQ2 ***Group Formation*** — Section 5.3.2								
0.00396	0.0628	0.0401	1.041	H2: BotRecruit (0/1)	0.0137	0.116	0.1385	1.149
0.0171	0.130	0.1121	1.119	H2: NomRecruit (0/1)	0.0374	0.190	-0.2213	0.802
0.00795	0.0888	-0.0850	0.918	H2: DeleteRecruit (0/1)	0.00635	0.0794	0.0562	1.058
0.00280	0.0528	0.3966	1.487 *	H2: KeepRecruit (0/1)	0.0357	0.185	-0.2448	0.783
RQ3 ***Experience*** — Section 5.3.3								
0.0756	0.134	0.0894	1.094	H3a: AfDNewcomers (%)	0.119	0.151	0.5278	1.695 **
0.0203	0.0711	0.4108	1.508 *	H3a: WPNewcomers (%)	0.0262	0.0754	2.3611	10.603 ***
0.591	0.270	0.0003	1.000	H3b: TenureDiversity	0.608	0.254	-0.0945	0.910 **
0.422	0.342	0.0003	1.000	H3b: TenureDiversitySq	0.434	0.324	-0.0091	0.991
RQ4 ***Administrative Bias*** — Section 5.3.4								
0.0544	0.227	0.0814	1.085	H4: AdmDeleteBias (0/1)	0.0260	0.159	-0.0100	0.990
0.193	0.394	-0.0826	0.921 +	H4: AdmKeepBias (0/1)	0.127	0.333	0.2051	1.228 **
			1132.29 ***	Likelihood Ratio				1142.14 ***

Table 5.1: Descriptive statistics of variables and results of logistic regression predicting flawed decisions. *Negative* β values and odds ratios (OR) *below* 1 indicate variables associated with better decision quality. (*** $p < .001$, ** $p < .01$, * $p < .05$, + $p < .1$)

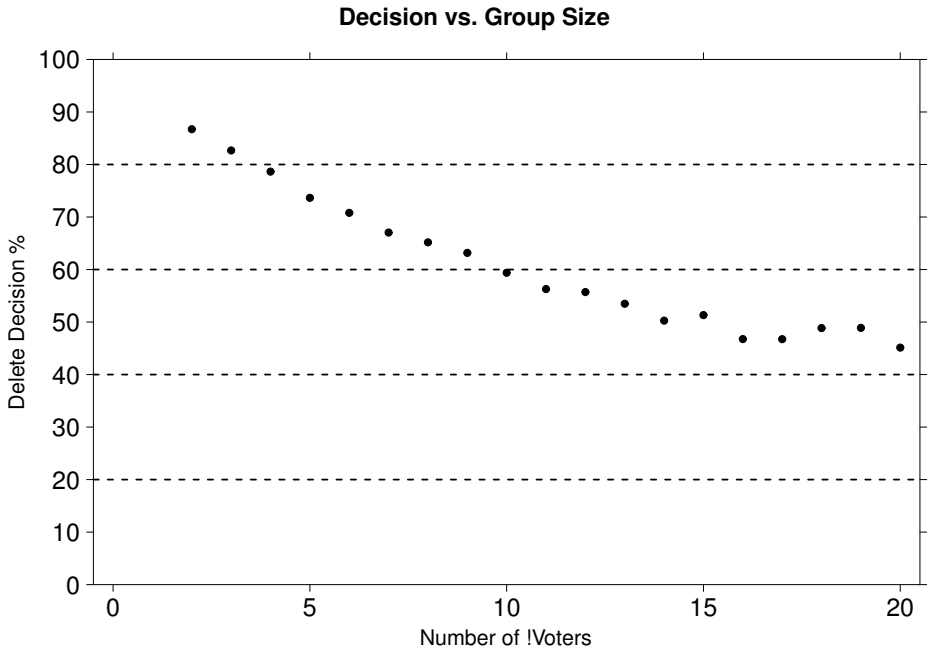

Figure 5.2: Relationship between AfD decision and group size.

We measured the effect of group size on decision quality by introducing *GroupSize*, a normalized variable containing the number of !voters in the AfD. To test for non-linear effects, we also added the quadratic term *GroupSizeSq*.

Model 1 and 2 both show that group size and its quadratic term are significantly associated with whether a decision is reversed. Figure 5.4 depicts the relationship described by both models. The plots are similar to one another and show that decisions made by small groups are more likely to be reversed than those made by larger groups, regardless of which decision was originally made. The plots flatten out toward the right, suggesting that there is little benefit from increases in size once a group is moderately sized.

5.3.2 RQ2: Group Formation

Wikipedia's AfD decision-making groups are self-formed, which, as discussed earlier, carries a risk of biased recruitment and membership. Also, group members could, in principle, strategically choose who to recruit in an deliberate attempt to influence the decision-making process. This is particularly true when group sizes are small and the addition of one or two people could

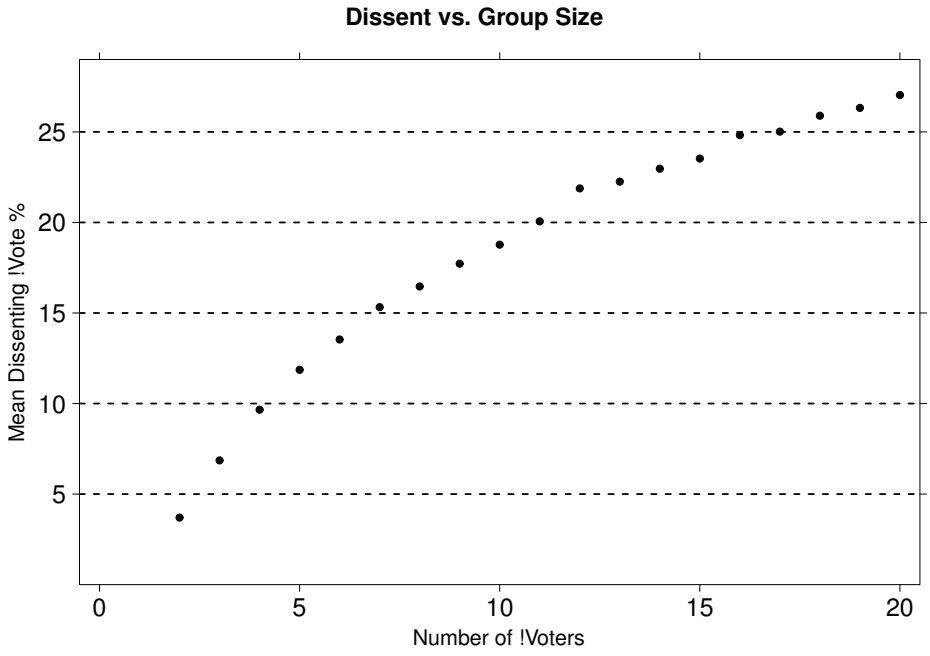

Figure 5.3: Relationship between group dissent and size in AfDs.

sway the perceived consensus, which is often the case with Wikipedia AfD discussions (recall that the median !voter count is four). To help avoid this sort of manipulation, Wikipedia's norms and policies only allow a limited form of direct recruitment. For AfDs, they permit neutral recruitment from two groups: the nominated article's primary contributors, and members of relevant WikiProjects (work groups that focus on particular topics).[6]

However, we observe that this policy itself has a form of bias. The permitted groups are comprised of people who likely have an interest in the nominated article, either because they helped write the article, or they have an interest in the article's topic area. Therefore, they may be predisposed to resist efforts to delete the article. There is valid reasoning for the policy though: because these groups' members are likely those who know the most about the article's topic, they are the most able to help make a well-informed decision about the article, and may be able to address any deficiencies in the article that caused its nomination for deletion.

That said, it remains the case that Wikipedia's recruitment policy contains bias. Also, the

[6]http://en.wikipedia.org/wiki/WP:AFDHOWTO

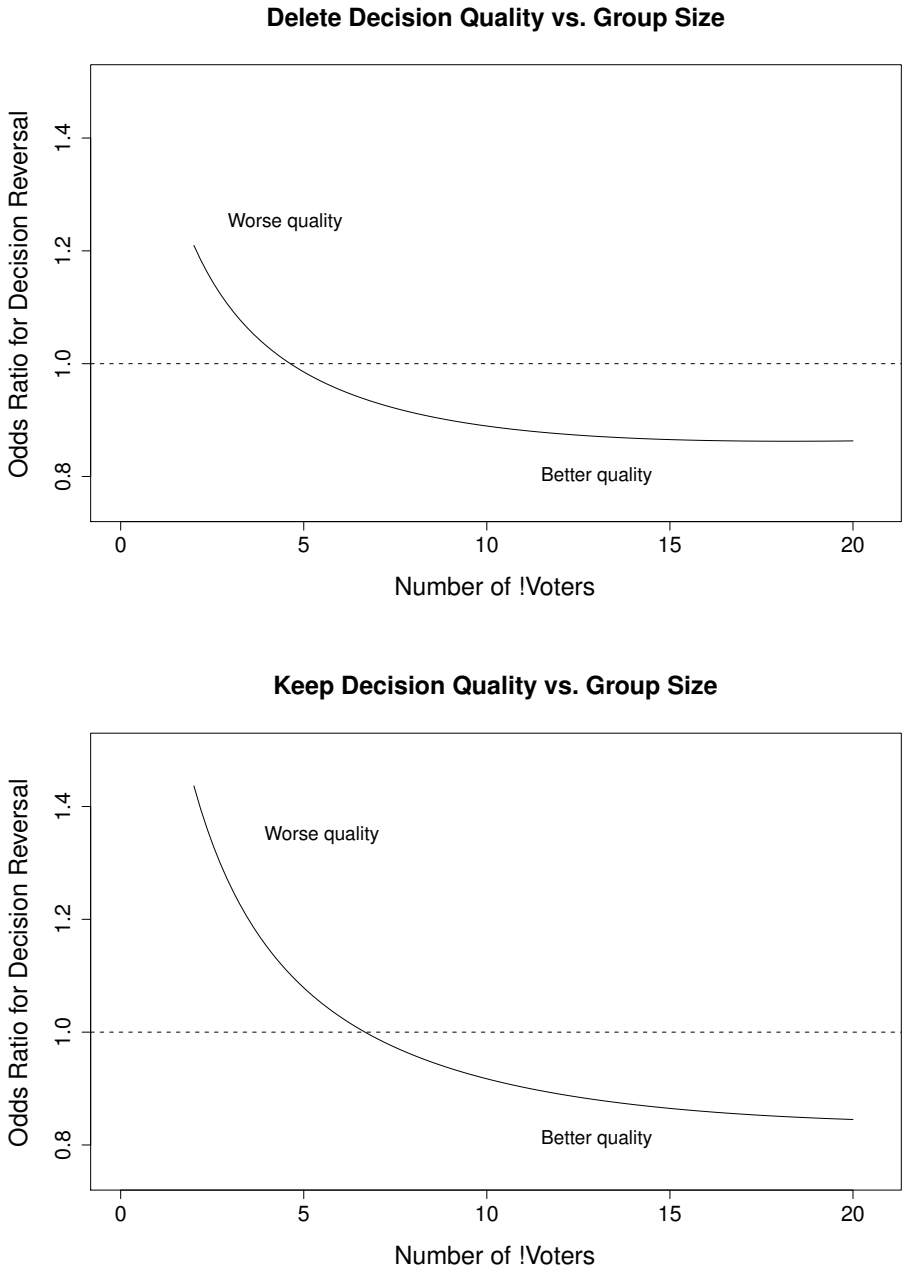

Figure 5.4: Effect of group size on decision quality (H1). Lower odds ratios indicate better decision quality.

policy may not be strictly enforced. There is no automated means of detecting improper recruitment (especially for recruitment that occurs outside Wikipedia), and manually investigating every participant is labor-intensive and draconian. We hypothesize the following about AfD decision-making groups that are formed through recruitment:

H2 *Recruit-Worse*: Groups formed through recruitment make worse decisions than naturally formed groups.

To test this hypotheses, we first need to observe and measure recruitment to AfD discussions. On Wikipedia, the typical method to communicate with a user is through *User Talk* pages, which are wiki pages associated with user accounts. So, to recruit a user to join a discussion, one would edit that user's *User Talk* page and write a message inviting him or her to the discussion. We detect cases of successful recruitment to AfD discussions by processing the metadata dump described earlier, looking for instances of the following sequence of events.

1. User *A* participates in AfD discussion *D*, either by nominating an article for deletion (thereby starting the discussion), or by expressing a !vote.
2. Within one hour or ten edits (whichever is sooner) of (1), user *A* edits user *B*'s User Talk page.
3. Within two days of (2), user *B* !votes in discussion *D*.

When we find such a sequence of events, we say that user *A* has successfully recruited user *B* to participate in AfD discussion *D*. We note that this is not definitive evidence of recruitment since it is possible that *A*'s message to *B* is unrelated to the AfD discussion, and that this sequence occurred due to coincidence. However, we believe this approach works reasonably well in practice, and is easily automated. Furthermore, we also apply the following heuristics to help filter out common false positives that we observed while performing manual spot-checks of the data.

First, if *A* or *B* had edited each others' User Talk page in the three days before the supposed recruitment message, then we did not consider the instance to be recruitment because *A* and *B* were probably in the midst of an unrelated conversation.

Second, if *B* was active in more than three other AfD discussions in the three days before or one hour after his or her !vote in *D*, then we did not consider the instance to be recruitment. We observed that active AfD participants tend to also be active elsewhere in Wikipedia. They receive many messages on their User Talk pages, including ones from other active Wikipedians

Recruiter	AfD Nom.	Delete !Voter	Keep !Voter	Bot
% of Recruits who !Voted Delete	34.6%	60.7%	15.1%	20.3%

Table 5.2: Summary of AfD group recruitment showing how recruited participants' !votes differ depending on who recruited them. Wikipedia-wide, 62% of AfD !votes are for deletion.

who are also frequent AfD participants. However, the messages often are not AfD recruitment messages, but are thank yous, warnings, feedback, or commentary regarding various topics that the users were involved with.

We also found that there have been two bots (computer programs that edit Wikipedia)—*BJBot* and *Jayden54Bot*—that automatically notified article editors about AfD discussions and recruited them per the established policy. These bots performed AfD notifications for several months, and offer us an opportunity to study the effect of recruitment that is purely policy driven. We use a process like the one described above to detect successful instances of bot-initiated recruitment: if a recruitment bot edited a user's talk page, and that user !voted in an AfD within two days, then we consider that user to have been recruited by the bot.

Using the above processes, we identified 8,464 instances of successful recruiting. Table 5.2 shows a summary of who did the recruiting, and how their recruits !voted. We see large differences in !voting behavior, which suggests that there is bias in who people choose to recruit. (From these data we cannot tell whether the bias is an intentional effort to influence consensus, or the result of social network homophily [103].) Participants recruited by delete !voters were about four times more likely to support deletion than those recruited by keep !voters. The participants that bots recruited also appear unlikely to support deletion, which reflects the inherent policy bias that we observed earlier.

To see what effect participant recruitment has on decision quality, we introduce four binary variables: *BotRecruit*, *NomRecruit*, *DeleteRecruit*, and *KeepRecruit*. These variables indicate whether a bot, the AfD nominator, a delete !voter, or a keep !voter successfully recruited somebody to the group, respectively.

Looking back to table 5.1, we find that regardless of the decision, none of the first three variables has a statistically significant effect. On the other hand, when a keep !voter recruited someone to the discussion, we see a significant effect: delete decisions are more likely to be reversed. We offer two possible explanations: the first is that recruitment by keep !voters, biased as it may appear, is a sign of positive community interest, and suggests that the article should be

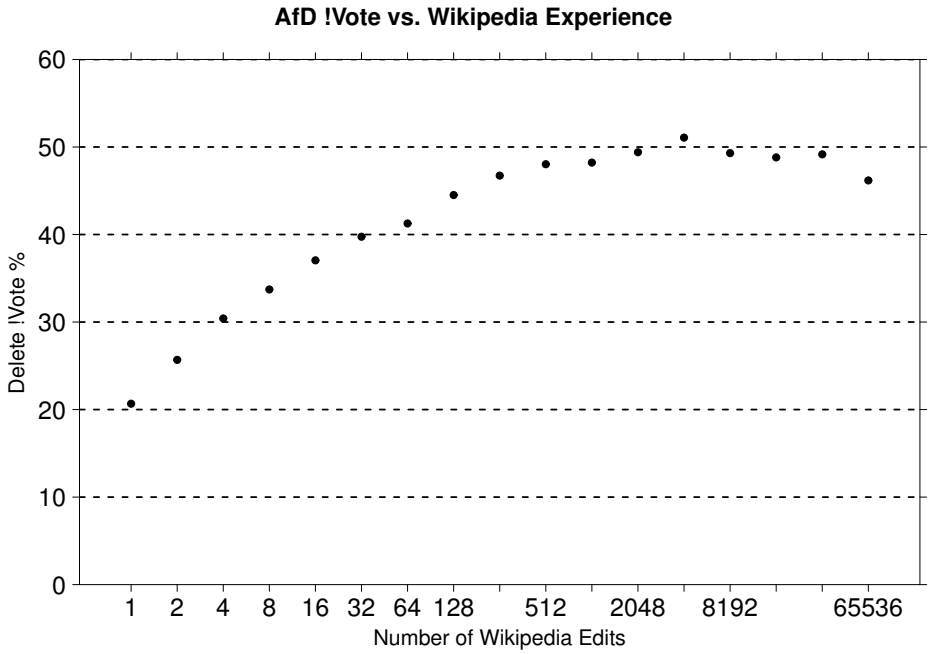

Figure 5.5: Relationship between AfD !vote and Wikipedia experience at time of !vote, computed on a per-user basis.

kept. If the AfD group decides otherwise and deletes the article, then decision quality suffers. An alternative explanation is that keep !voter recruitment is a sign of activism among those who prefer to keep the article. These proponents may be especially persistent in maintaining the article's existence in Wikipedia, even if it requires working to reverse a delete decision.

5.3.3 RQ3: Experience

Next, we turn to our research question regarding the role of participant experience and tenure diversity in decision quality. Our exploration of Wikipedia AfD discussions suggests that there are fundamental differences between newcomer and oldtimer behavior. As shown in figure 5.5, users who have little Wikipedia editing experience are far less likely than experienced users to !vote for deleting an article. Newcomers' knowledge and interpretation of Wikipedia's article policies are evidently different from that of oldtimers, perhaps due to lapses in newcomer socialization. This suggests that newcomer participation in AfD discussions may adversely affect decision quality.

By contrast, diversity theory says that increased tenure diversity can lead to better group outcomes. A study by Chen, et al. showed that moderate tenure diversity in Wikipedia's WikiProjects (topical work groups) is beneficial to productivity and retention [22], and we believe the effect will extend to decision quality. We hypothesize:

H3a *Newcomers-Worse*: Groups with more newcomers make worse decisions than groups with fewer newcomers.

H3b *Diversity-Moderate*: Groups with moderate tenure diversity make better decisions than groups with high or low tenure diversity.

To test H3a, we introduce two measures of newcomer participation to our model: percentage of participants in each discussion who had 15 or fewer Wikipedia edits (*WPNewcomers*), and percentage of participants who were not new to Wikipedia, but had !voted in five or fewer AfDs (*AfDNewcomers*). To test H3b, we use the normalized tenure diversity of the participants who !voted in each AfD discussion (*TenureDiversity*), and its quadratic term (*TenureDiversitySq*). Our definition of tenure diversity is identical to the one used by in [22]: the coefficient of variation of the number of days since each group member's first Wikipedia edit. The coefficient of variation is a widely used measure of tenure diversity in past research [10].

We start by looking at H3a. Both models in table 5.1 show that decisions made by groups that have Wikipedia newcomers are significantly more likely to be reversed. Model 2 shows that participation by AfD newcomers also leads to more reversals when they are involved in making a keep decision. Proceeding to H3b, we turn to the tenure diversity measures. We find that tenure diversity has no effect when the decision is to delete. However, diversity has a significant effect when the decision is to keep. A plot of the effect on odds ratio is shown in figure 5.6, and indicates that decision quality improves with tenure diversity. It is unclear why tenure diversity's effect appears dependent on the decision outcome.

5.3.4 RQ4: Administrative Bias

Finally, we look at RQ4, which is about the effect of administrative bias on decision quality. In Wikipedia AfDs, the administrator who closes the discussion is responsible for identifying what the community has decided to do. To do this, the administrator applies guidelines that describe how to interpret the discussion and determine whether a rough consensus was reached.[7]

[7] http://en.wikipedia.org/wiki/Wikipedia:DGFA

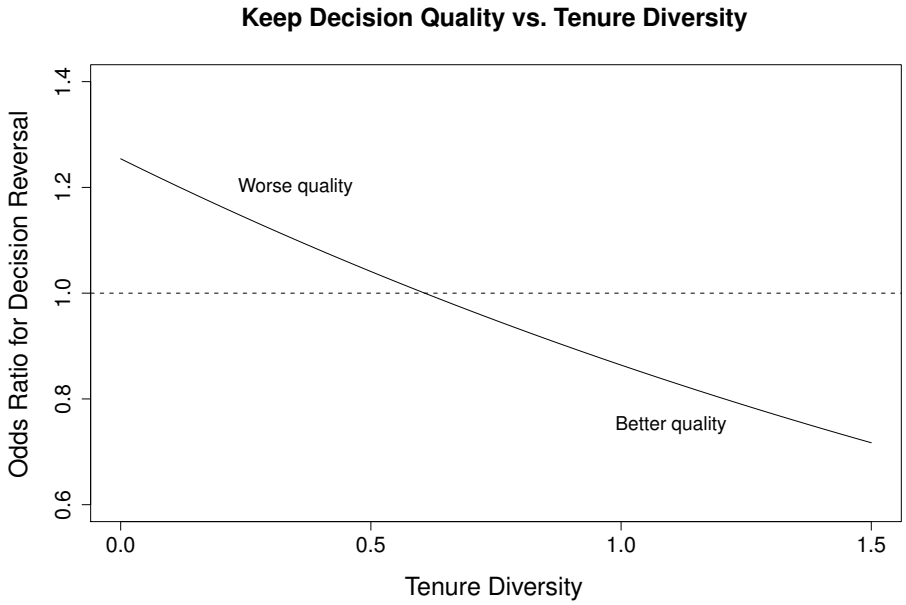

Figure 5.6: Effect of tenure diversity on decision quality (H3b). Lower odds ratios indicate better decision quality.

The guidelines allow for some subjectivity. They require that the administrator to use his or her "best judgment," but to "be as impartial as is possible for a fallible human." Because hundreds of administrators volunteer to close AfD discussions, and because determining consensus requires human judgment, there exists opportunity for administrative bias to affect the AfD decision-making process. We expect that groups with biased administrators will yield poorer decision quality than those with unbiased ones.

H4 ***Biased-Admin-Worse*****:** Groups with biased administration make worse decisions than groups with neutral administration.

We measure administrative bias by looking at how different administrators make consensus calls in AfDs that have similar !vote breakdowns, and comparing their behavior to a Wikipedia-wide statistic. To provide some intuition, let us consider AfDs in which three participants !voted keep and four !voted delete. Historically, 56% of such AfDs have resulted in a decision to delete the nominated article. Now, suppose that Fred, an administrator, has closed ten of these AfDs, and that he determined there was a consensus to delete the article in two of them, or 20%.

Since 20% is much less than 56%, the evidence suggests that Fred is biased away from delete outcomes and towards keep outcomes. (Administrators may also be biased toward certain topics or articles, but we do not consider such biases in the current work.)

To turn this approach into a bias metric, we first divide all AfDs into 11 groups based on each AfD's !vote breakdown, measured as percentage of !votes in favor of deletion. We ignore !votes that are for outcomes other than keep and delete (96% of !votes are to keep or delete the article). The first group contains AfDs with fewer than 5% delete !votes. The second group contains AfDs with at least 5% but fewer than 15% delete !votes. The third group contains AfDs with at least 15% but fewer than 25% delete !votes, and so on. The final group contains AfDs with at least 95% delete !votes.

Now, we can define a relative bias measure for an administrator A by summing the differences between his or her consensus calls and the Wikipedia-wide ones for each group. Because administrators may have little or no data in some groups of AfDs, we apply a form of Bayesian smoothing that is based on Wikipedia-wide statistics.

$$bias_A = \sum_{i=1}^{11} \left(\frac{C * p_i + numdel_{A,i}}{C + numafds_{A,i}} - p_i \right) \tag{5.1}$$

Here, p_i is the Wikipedia-wide percentage of AfDs in group i that were closed with a delete decision, $numafds_{A,i}$ is the number of AfDs in group i that A closed, and $numdel_{A,i}$ is the number that A closed with a delete decision. C is a tunable parameter used for smoothing. For our analysis, we set $C = 3$. Using this definition, we compute a bias for all administrators who have closed at least ten AfD discussions. Such administrators collectively closed 91% of the AfDs in our data set.

If the bias measure is positive, then the administrator tends to close discussions with a delete decision more often than average, given some !vote breakdown. Similarly, if the bias measure is negative, he or she is more likely to close discussions with a keep decision. The measure's magnitude represents the strength of the apparent bias. To make this measure easier to interpret, we normalize these values to standard scores. For example, a measure of -1.4 indicates that the administrator's apparent bias is 1.4 standard deviations from the average, and that the bias is in the keep direction.

To illustrate the differences that we see among administrators, figure 5.7 shows the consensus calls profiles for two administrators who appear diametrically biased with bias measures

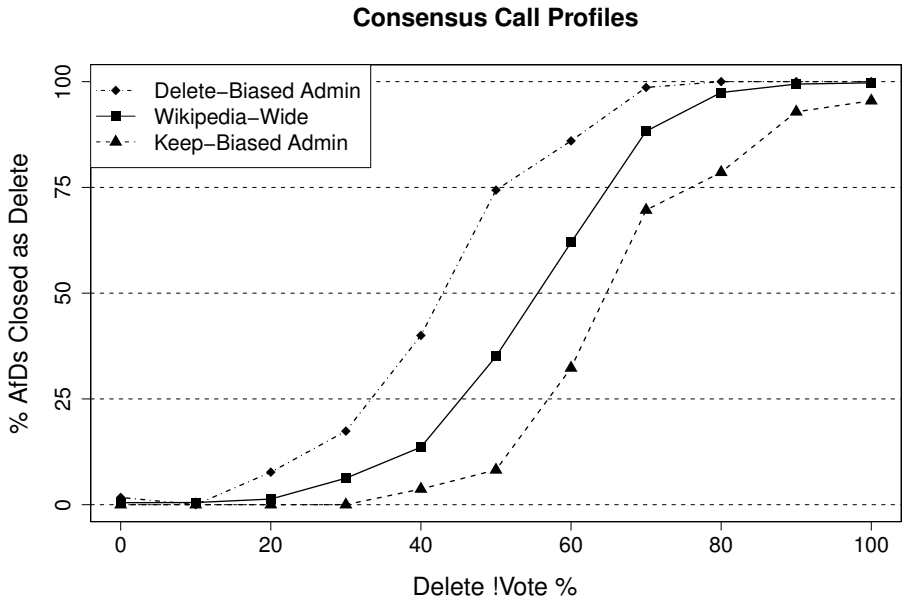

Figure 5.7: Relationship between !vote breakdown and decision. The two dotted lines show the consensus call profiles for two diametrically biased administrators with high-magnitude bias measures (-2.92 and +2.58).

of -2.92 and +2.58. For comparison, we have included the Wikipedia-wide average consensus call profile. These plots show the likelihood that an AfD results in a delete decision for each of the 11 groups. Indeed, there appears to be substantial variation in how different administrators determine consensus.

To elucidate this point, we note the large magnitude of the difference in the center data point value (50% Delete !Vote) for the two administrators. In these evenly-split AfDs, the delete-biased administrator found a consensus to delete the article in about 75% of cases, while the keep-biased administrator only found such a consensus in less than 10% of cases!

Recall that Wikipedia's policy regarding decision making is that consensus should *not* be determined according to vote counts. However, we observe that the plots in figure 5.7 resemble logistic curves, suggesting that Wikipedia's rough consensus process approximates a vote, but permits administrators to, at their discretion, accept a compelling minority opinion as the decision. It is interesting to note that on average, there is a slight tilt toward keep decisions; the solid line in figure 5.7 shows that it requires more than 50% of !voters in an AfD to favor

deletion before the likelihood of a deletion decision reaches 50%.

These measurable aggregate-level differences among administrators in when they accept minority opinions are suggestive of a bias in which some administrators are systematically discounting certain opinions, perhaps subconsciously. To determine the effect of this apparent bias in AfD discussions, we introduce a categorical variable that denotes whether the closing administrator is keep-biased (bias of less than -2), delete-biased (bias of greater than +2), or neutral (bias of -2 to +2, or uncomputable[8]). We encode this as two dummy-coded variables, *AdmKeepBias* and *AdmDeleteBias*, with neutral administrators as the reference group.

Model 1 indicates that when keep-biased administrators are involved in delete decisions, there is a marginally significant *decrease* in reversals ($p = 0.0594$). If an administrator makes a consensus call that is contrary to his or her own bias, the community's arguments were likely strong enough to overcome that bias, thus leading to an increase in decision quality. On the other hand, model 2 shows that when these keep-biased administrators make the call to keep an article, their decisions are reversed more often. We see no significant effect on decision quality when delete-biased administrators are involved in decision making, but the models show weak trends consistent with the results for keep-biased administrators.

5.4 Discussion

In each part of section 5.3, we looked at the results of our analysis as they relate to one of our research questions. Now, we will step through the research questions and discuss our findings, as well as the implications for social production community design.

RQ1: Group Size. We find support for **H1 *Bigger-Better***. Online decision-making processes that involve too few people are at higher risk of making low quality decisions. Larger groups make better decisions, but with rapidly diminishing returns.

Our findings lead us to two design suggestions. First, encourage more users to participate in collaborative decision-making activities. The increased group sizes can improve decision quality. However, since we see evidence of diminishing returns, it may be beneficial to steer users toward underpopulated areas instead of toward areas that are already crowded. Second, be wary of decisions that are made by groups that are very small, as they may be suspect. Scrutinize the decisions carefully, and consider delaying the decision to find additional participants if

[8]Recall that we do not compute a bias metric for administrators who closed fewer than 10 AfDs.

there does not appear to be a sufficient quorum. In communities where decisions are made in a structured manner, it may be possible to automate both suggestions through intelligent task routing techniques [29].

RQ2: Group Formation. In our exploratory analysis, we found strong evidence of biased recruitment to AfDs. People appeared to seek out like-minded peers. Despite the biases, our results only show limited support for **H2** ***Recruit-Worse***—decision quality is unaffected by most forms of recruitment that we studied. More work is needed to understand why this is the case. Perhaps Wikipedians are aware of these biases and are able to adjust accordingly.

However, our results do shed some light on the complexity that exists when decision-making groups are self-formed. Designers should carefully consider possible biases when constructing policy about how to attract participants (e.g., Wikipedia's policy bias as described in section 5.3.2). To reduce the amount of human effort required for recruitment, and to help avoid biases from selective recruitment, communities may wish to consider automating outreach strategies (e.g., Wikipedia's AfD notification bots). Additionally, it may be possible to construct automated tools that look for signs of biased group formation, and to carefully scrutinize decisions made by such groups.

RQ3: Experience. We find partial support for **H3a** ***Newcomers-Worse*** and **H3b** ***Diversity-Moderate***. Newcomer participation is detrimental to decision quality, while high tenure diversity is beneficial in some cases. The latter finding partially disagrees with what diversity theory predicts, and suggests AfD groups do not suffer from the negative social categorization effects of high tenure diversity [22, 156]. A possible explanation for this is that AfD groups are ephemeral and highly task-oriented. Thus, conflict between newcomers and oldtimers might not appear in an AfD discussion, but may negatively affect their interactions elsewhere on Wikipedia.

In light of these findings, we recommend that SPCs encourage all users (including newcomers) to participate in group decision-making processes, but with a focus on socializing newcomers to help them understand issues related to the decisions at hand. Automated tools could also watch for and draw attention to situations that lack sufficient newcomer or oldtimer presence. Taking steps to increase group diversity in these cases may provide improved decision quality as well as provide opportunities for newcomer socialization.

Looking back at figure 5.5, we are intrigued by the extent to which newcomers and oldtimers apparently differ in their opinions in AfDs. We speculate that in cases like this where

newcomers tend to disagree with oldtimers, it could be beneficial to give serious thought to newcomer input and consider ways to integrate their ideas into community norms. Rebuking newcomers' opinions in favor of existing group ideals could alienate and drive away newcomers, which may be harmful in the long term. In Wikipedia's case, this may be a contributor to observed slowdowns in growth [146]. Established norms can certainly be difficult to change, but acceptance of new ideas may be necessary to keep a community sustainable.

RQ4: Administrative Bias. We find support for **H4** ***Biased-Admin-Worse***. The presence of biased leaders can lead to worse decision quality when they are involved in decisions that agree with their bias. However, when they are involved in decisions contrary to their bias, we find evidence that decision quality improves. These findings lead us to two design recommendations.

First, builders and maintainers should be conscious of the possibility of administrative bias. There should be a process for the community to challenge questionable decisions and inconsistent administration. For example, Wikipedia AfD decisions can be appealed through a process called Deletion Review. Such processes will help the community identify potentially problematic individuals, and may lead to improved decision quality.

Second, consider automated mechanisms that draw attention to contentious cases, especially if they involve an administrator who has acted in concordance with a history of apparent bias. By introducing additional community discussion or analysis by a secondary administrator, it may be possible to reduce the negative impact of any biases that might exist and effect better decisions.

Summary. In this chapter, we have explored how four group composition factors influence decision quality in a large online SPC. Our findings are summarized in table 5.3. Earlier in this section, we provided discussion and recommendations that we hope will inform the design of more effective decision-making processes and tools. While not all of our findings are definitive, we believe they raise interesting questions for SPCs (e.g., how should a group fairly canvass the community for useful input, or address inconsistencies in administration?), and they point the way toward future work.

Our work focused on one class of content decisions made on Wikipedia, and thus, our results may not generalize to other types of decisions or other communities. Further work is necessary to test our results in different environments. Nonetheless, we believe that the decisions we studied—ones of content relevance and appropriateness to the community—are representative

Hypothesis	Result	Description
H1 ***Bigger-Better***	Supported	Larger groups make better decisions, but with diminishing returns
H2 ***Recruit-Worse***	Mixed	Biased recruitment leads to worse decisions under some circumstances
H3a ***Newcomers-Worse***	Supported	Newcomer participation yields worse decision quality
H3b ***Diversity-Moderate***	Mixed	Diverse groups may make better decisions; no social categorization effects were observed
H4 ***Biased-Admin-Worse***	Supported	Worse decisions in some cases if decision agrees with administrator's bias Better decisions in some cases if decision is contrary to administrator's bias

Table 5.3: A summary of our findings.

of curation decisions that SPCs typically face, and that our contributions here will help drive the development and evolution of future systems.

Chapter 6

Effects of Gender Imbalance Among Wikipedia Editors

6.1 Introduction

In the previous chapter, we demonstrated how various skews in small-group composition are associated with differences in the quality of individual curation decisions. Now, we turn to the bigger picture – are there skews in the makeup of entire populations in social production communities? If so, do these skews induce far-reaching system-wide effects, especially in how collaborative curation is affected? Here, we present a detailed study of one broad and large-scale imbalance in an SPC's population, and its effects on how the SPC's repository of information has been curated.

A January 2011 New York Times article by journalist Noam Cohen described an interesting but perhaps troubling phenomenon affecting Wikipedia: a wide gender gap amongst its community of volunteer editors [25]. In the article, Cohen observes that just 13% of Wikipedia's contributors are female, according to a 2009 Wikimedia Foundation survey. Furthermore, he suggests that this disparity has led to deficiencies in Wikipedia's coverage of "female" topics, as evidenced by a series of anecdotal examples (e.g., Wikipedia's coverage of stereotypically-female topics like friendship bracelets or "Sex and the City" pales in comparison to that of stereotypically-male topics like toy soldiers or "The Sopranos").

The Wikimedia Foundation has acknowledged the presence of a gender gap, and has established a goal of increasing the female share in editors to 25% by 2015. While ambitious, such an

accomplishment may be attainable. In *Unlocking the Clubhouse: Women in Computing* [101], Margolis and Fisher describe a series of studies and educational reforms that helped Carnegie Mellon University address a wide gender gap in their undergraduate Computer Science program. Over the course of five years, the percentage of women entering the program rose from 7% in 1995 to 42% in 2000.

Before attempting to address any imbalance that might exist in Wikipedia, it is important to first understand the nature of that imbalance. We conduct a quantitative exploration of gender imbalance in English Wikipedia's volunteer editor population, and the effects it has had on the encyclopedia. Cohen's article presents a compelling argument, but we believe there is need for more rigorous analysis that expands on the reported survey results and anecdotal evidence. We believe our work represents a crucial next step in understanding what is happening and deciding what can be done to address it.

6.1.1 Related Work

Research from the volunteering literature and the technology adoption literature offer reasons not to expect a large gender gap among Wikipedia's editors. Taniguchi finds that females are more likely to volunteer than males, and that females do more volunteer work than males [151]. In [170], Wilson cites four underlying reasons for females' increased volunteerism: they exhibit greater empathy and altruism, they place more value in helping others, they perceive a gender-specific norm that they should take care of others, and they view volunteering as part of their "social life." Overall, these findings suggest that females may be more likely to volunteer their time to edit Wikipedia, though they may edit less if they lack a social connection to the Wikipedia community.

The technology adoption literature suggests that females may lag behind males in adopting new technology. Broadly, Venkatesh et al. find that females are less likely to adopt new technologies than males, and that females are more heavily influenced by social norms related to a technology and the perceived difficulty of the new technology [157]. However, studies of gender differences in adoption of the Internet and social media offer more encouraging findings. Periodic Pew Research Center surveys of the general populace show that Internet usage between 2000 to 2004 was skewed toward males, but that the gap has since disappeared [117]. Females are now *more* likely than males to participate in some social media sites such as Facebook or MySpace [155]. In addition, females are more likely to tweet (10% of females, 7% of males),

and teenage girls are more likely to blog (25% of girls, 15% of boys) [143, 93]. Even online gaming, which is traditionally seen as a male-dominated activity [175], shows signs of a sea change; market research surveys indicate that females and males are on par with each other in online social gaming [77, 150].

Together, these observations suggest that Wikipedia, a community of volunteers collaborating to build a vast educational resource, ought to have a reasonable gender balance. However, multiple studies have indicated an apparent disparity. Lim's surveys of college students find that while all respondents had used Wikipedia, females visited it less frequently and perceived it to be of lower quality than males did [96]. A 2011 Pew Research Center survey finds a small gender gap in readership (50% of female Internet users, versus 56% of male ones) [177]. The Wikimedia Foundation commissioned a survey of Wikipedia users in 2009, and its results show a large gap among readers (75% male, 25% female), and an even larger gap among editors (87% male, 13% female) [48]. However, because users self-selected to participate in the survey, the report authors acknowledge that it is "hard to evaluate whether the shares we found in our survey are representative." Furthermore, 75% of the users who took the survey were using a non-English Wikipedia. Thus, it is uncertain what the gender gap is in English Wikipedia.

6.1.2 Contributions

Our work seeks to explore more carefully the state of gender imbalance among the English Wikipedia's volunteer editors. We extend existing research on Wikipedia and gender in three key ways. First, we conduct a high-level study of gender and editing behavior in order to measure and characterize the editor gender gap. Second, we explore how the imbalance affects Wikipedia's content and community. Finally, we analyze the role of gender in conflict among Wikipedia's editors to help understand why an imbalance might exist.

6.2 Research Questions

We begin our exploration of Wikipedia's gender gap by posing three overarching research questions and seven hypotheses.

6.2.1 RQ1: Gap-Overall

What is the extent of Wikipedia's gender gap, and how has it changed over time?

We are interested in measuring Wikipedia's gender gap and determining whether the imbalance is growing or shrinking. Based on the survey results presented in [48], we hypothesize that Wikipedia does have a wide gender gap. Note that the remainder of our hypotheses and research questions provisionally assume that this hypothesis will be supported.

H1a *Gap-Exists*: Wikipedia has a substantial editor gender gap.

Periodic surveys of the general populace indicate that there was a modest gender gap in Internet use in the early 2000s, but that it has been shrinking steadily [117]. We hypothesize that a similar trend has been taking place in Wikipedia.

H1b *Gap-Shrinking*: Wikipedia's gender gap is shrinking.

6.2.2 RQ2: Gap-Matters

How is Wikipedia affected by the gender gap?

In this research question, we wish to explore whether the gender gap is causing some parts of Wikipedia to receive less attention than other parts. Our next hypothesis is inspired by Noam Cohen's New York Times article [25], which provides anecdotal evidence of a large disparity in Wikipedia's depth of coverage for "female" topics as compared that of "male" topics. We systematically study this phenomenon in large-scale data-driven ways that do not depend on invocation of gender stereotypes to determine which topics are "female" or "male." Formally:

H2a *F-Coverage-Worse*: Coverage of topics with particular interest to females is inferior to topics with particular interest to males.

We further hypothesize that due to gender differences in extraversion, empathy, and altruism [38, 170], females will tend to be more active than males in social- or community-oriented areas of Wikipedia that offer increased interaction with other editors and opportunity to build interpersonal relationships. If this hypothesis is supported, addressing the gender gap might lead to a healthier community in which there are more resources available for community-oriented tasks like helping new editors and organizing editor efforts.

H2b *F-Social*: Females are more likely to be involved in social- and community-oriented areas of Wikipedia.

6.2.3 RQ3: Gender-Conflict

What gender differences exist in conflicts in Wikipedia, and how do those differences relate to the gender gap?

In our final research question, we look at gender differences in conflict among Wikipedia's editors in order to learn about how conflict might be contributing to the gender gap. Prior research finds that conflict has been a growing problem for Wikipedia, consuming increasing amounts of editor effort [85]. Studies on gender and personality have shown that females tend to have more agreeable and less aggressive personalities [38, 13], which suggests that they may tend to avoid conflict. Therefore, a possible explanation for the gender gap is that females find conflict among Wikipedia editors to be distasteful and unappealing, and simply choose to not edit Wikipedia as a result. As a partial test of this explanation, we hypothesize that Wikipedia's existing female editors tend to do their work in less controversial areas.

H3a *F-Uncontentious*: Females tend to avoid controversial or contentious articles.

Females who do decide to edit Wikipedia may find it difficult to make contributions that are accepted by the community. Drawing upon years of research in gender and computer-mediated communication, Herring finds that "gender differences in on-line communication tend to disfavor women" and that females who participate in mixed-gender online environments tend to be marginalized [71]. Thus, we believe there may be a systemic bias against females that cause their edits to be more likely to be reverted (undone by another editor), particularly early on in their Wikipedia tenure. Previous work has shown that being reverted as a newcomer is particularly demotivating [58]. Furthermore, since females may prefer to avoid conflict, we believe they are more likely than males to lose interest and leave Wikipedia if their early contributions are reverted.

H3b *F-Reverted-More*: Female editors are more likely to have their early edits reverted.

H3c *F-Reverted-Leave*: Female editors are more likely to stop editing and leave Wikipedia when being reverted as newcomers.

6.3 Data

To test our hypotheses and answer our research questions, we perform a variety of quantitative statistical analyses on publicly-available English Wikipedia data. The majority of our data is either from the January 2011 data dump[1] or from the Wikipedia website itself (collected during February and March 2011 via the public API or screen-scraping). We also drew upon several other sources of data during our analyses, including a data set derived from an older full-text Wikipedia dump released in January 2010. Because each of these additional sources of data is specific to one of our hypotheses, we defer discussion about these other sources until the relevant section.

6.3.1 Editor Gender Data

A key piece of information we need for our analyses is editor gender. In Wikipedia, there are several ways that an editor can publicly disclose his or her gender:

1. They can specify whether they are male or female in their account's preference settings. The gender setting is described as being "used for gender-correct addressing by the software," and is public information available from Wikipedia's API.
2. They can place a gender userbox on their User page to openly announce and display their gender in a de facto standard way.
3. They can mention their gender while describing themselves on their User page or during a discussion with fellow editors in one of Wikipedia's discussion areas.

The editor gender data used in our analysis includes users who disclosed their gender via preference setting (#1) and a subset of those who used a userbox (#2). Figures 6.1 and 6.2 show screen shots of these two disclosure methods. For #1, we queried the Wikipedia API to obtain the gender setting for all users with at least one edit, and for #2, we identified users who have either the *User:UBX/male* or *User:UBX/female* gender userbox displayed on their User page. There exist several other userboxes that can be used to denote one's gender (including ones for transgender and other alternative genders), but since they have been used by too few users to give statistically reliable data, we chose to not include them in our analysis.[2] We also chose to

[1] http://dumps.wikimedia.org/enwiki/20110115/

[2] For example, the *User:UBX/transgender* userbox appears on fewer than 20 Wikipedia editors' User pages.

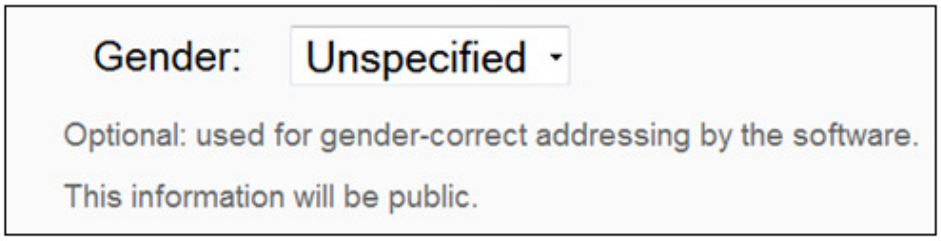

Figure 6.1: Gender preference setting in its default state.

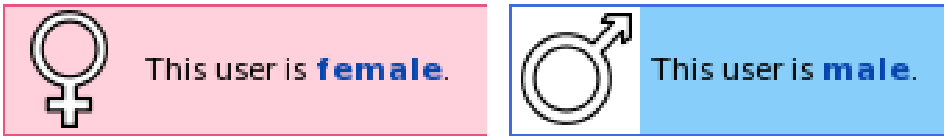

Figure 6.2: Gender userboxes used by Wikipedia editors to display and announce their gender.

not collect gender disclosures that occurred only via method #3 because doing so would require substantially more advanced natural language processing techniques.

We collected data for 106,698 users that have disclosed their gender using the user preference setting and 8,630 users who have done so using a userbox. Of these, 1,478 users specified their gender using both methods. The vast majority of these users specified the same gender in both places. However, two users have a different gender specified for each disclosure method, so we excluded them from our data set. In total, our editor gender data set contains 113,848 users who have collectively made over 67 million edits.

6.3.2 Assumptions and Limitations

This gender data is comprised of self-reports from self-selected users, and thus has limitations. The validity and generalizability of our results are subject to several assumptions about this data. First, we assume that users are mostly honest in reporting their gender. Second, we assume that users who do not report their gender behave similarly to those who do report their gender (which is just 2.8% of editors). Finally, we assume that self-report rates are similar between males and females at similar stages of their Wikipedia life-cycle. In our data, we find that self-report rates increase dramatically for dedicated editors: we have gender information for 6.5% of editors with at least ten edits, 14.1% of those with at least 100 edits, and 34.7% of those with at least

1,000 edits.

While necessary to enable our analysis, these assumptions are difficult to confirm. We do note that at a high level, our results are comparable to those obtained from the Wikimedia Foundation's recent survey of Wikipedia users [48], which provides limited support for our assumption about truthful self-reports. However, we are unable to provide evidence for the other assumptions. Unfortunately, this problem is fundamental for any Wikipedia research that depends on existing self-reported data. Most of the results in this research are subject to these assumptions and limitations.

6.4 Results and Analysis

We now step through each of our research questions and hypotheses, describing our methodology for testing each hypothesis, the data we used, and our results.

6.4.1 RQ1: Gap-Overall

H1a *Gap-Exists*: Wikipedia has a substantial editor gender gap.

To characterize the gender gap at a high level, we compared male and female editors using three broad metrics: editor count, edit count, and activity lifespan. We found that females comprised 16.1% of the 38,497 known-gender editors who started editing Wikipedia during 2009. This is indicative of a substantial gender gap in Wikipedia editors – males outnumber females by over 5 to 1. However, this is not the end of the story. Our other two metrics suggest that the gender gap is even deeper than indicated by simple editor count.

We found that despite females being 16.1% of the new editors in 2009, they only accounted for 9.0% of edits made by this cohort of editors. On average, a male editor made almost twice as many edits as a female editor. Figure 6.3 depicts the gender gap at various levels of edit count. We see that the gender gap widens when looking at editors with increasingly many edits, and does not appear to stabilize until the percentage of female editors drops to around 6% for editors making more than about 500 edits. This observation points to the possibility that females leave Wikipedia earlier than males in their editing tenures. Our third metric, activity lifespan, directly examines this possibility.

Activity lifespan measures how long a user is active in a system before leaving. We used

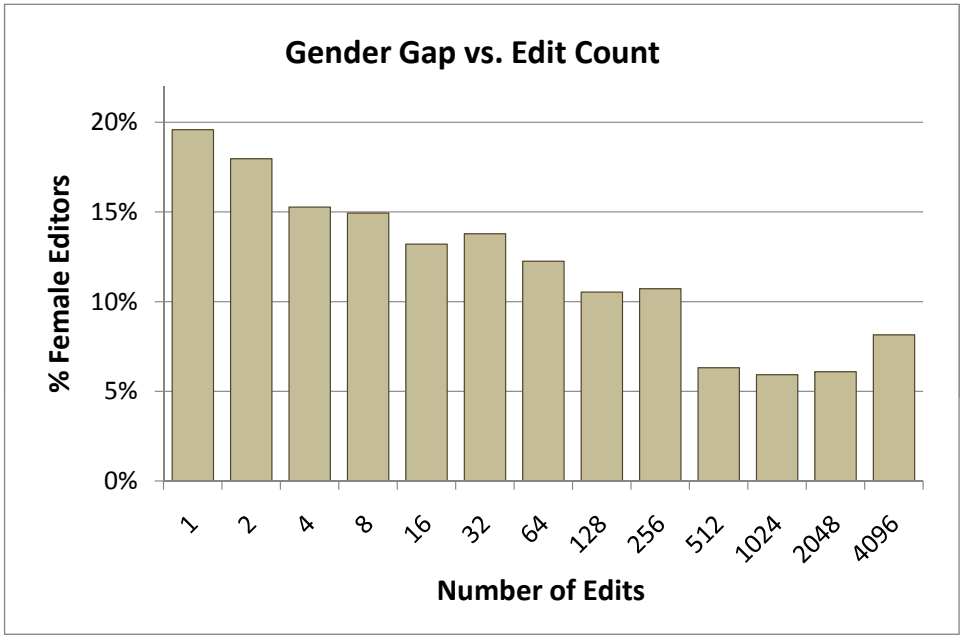

Figure 6.3: Wikipedia's gender gap as a function of editor activity for editors first editing Wikipedia during 2009. The gender gap is more pronounced when looking at high-activity editors.

a definition of activity lifespan similar to Yang et al.'s in their work studying user survival in social question-answering systems [174]. We considered an editor's "birth" to be his or her first edit date and a "death" to be a period of edit inactivity exceeding six months (the death is recorded as the beginning of the period). Using this notion of user birth and death, we applied standard survival analysis techniques [31]. Figure 6.4 shows the estimated male and female survival curves for users joining during 2009. We see a distinct difference between the two curves. Females stop editing Wikipedia sooner than males, and the ratio of males remaining to females remaining for this cohort increases steadily as time passes. Therefore, one of the factors contributing to Wikipedia's gender gap is a lower retention rate for female editors compared to male editors. We will return to exploring this survival data in greater detail in section 6.4.3.

H1b *Gap-Shrinking*: Wikipedia's gender gap is shrinking.

To examine how the overall gender gap has evolved over time, we looked for a trend in the gender breakdown of new editors joining Wikipedia each month. However, there are confounds in the data that made this a tricky task. The two methods for disclosing one's gender that we

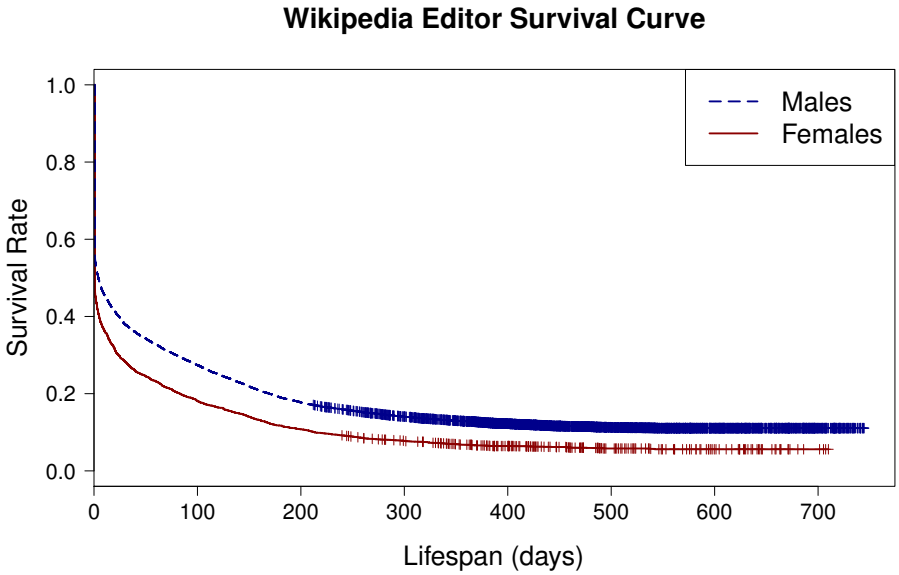

Figure 6.4: Activity lifespan of new female and male Wikipedia editors in 2009 expressed as Kaplan-Meier survival curves. The difference between the curves is statistically significant (logrank test, $p < 0.001$).

consider in our data (userboxes and preference setting) were introduced at different times, so simply looking at the trend over all of Wikipedia's existence is not a fair analysis. Users who started editing Wikipedia before a gender disclosure method was introduced could not specify their gender using that method until after its introduction date. The survival analysis presented above shows that males tend to have longer activity lifespans. Therefore, males who joined Wikipedia before a gender disclosure method was introduced are more likely than females to still be active once the method is made available (and thus, be able to specify a gender using it).

Due to this confound, we could only make valid comparisons for users joining Wikipedia after the introduction of a gender disclosure method, and only for the subset of users who used that disclosure method. The gender userboxes were introduced in December 2005, and the gender preference setting was introduced in January 2009. Figure 6.5 shows two charts, each depicting the gender gap over time for one of the gender disclosure methods. The trends in both charts are flat with essentially zero slope. Therefore, Wikipedia's gender gap appears to have remained approximately constant since December 2005, which is surprising given that other

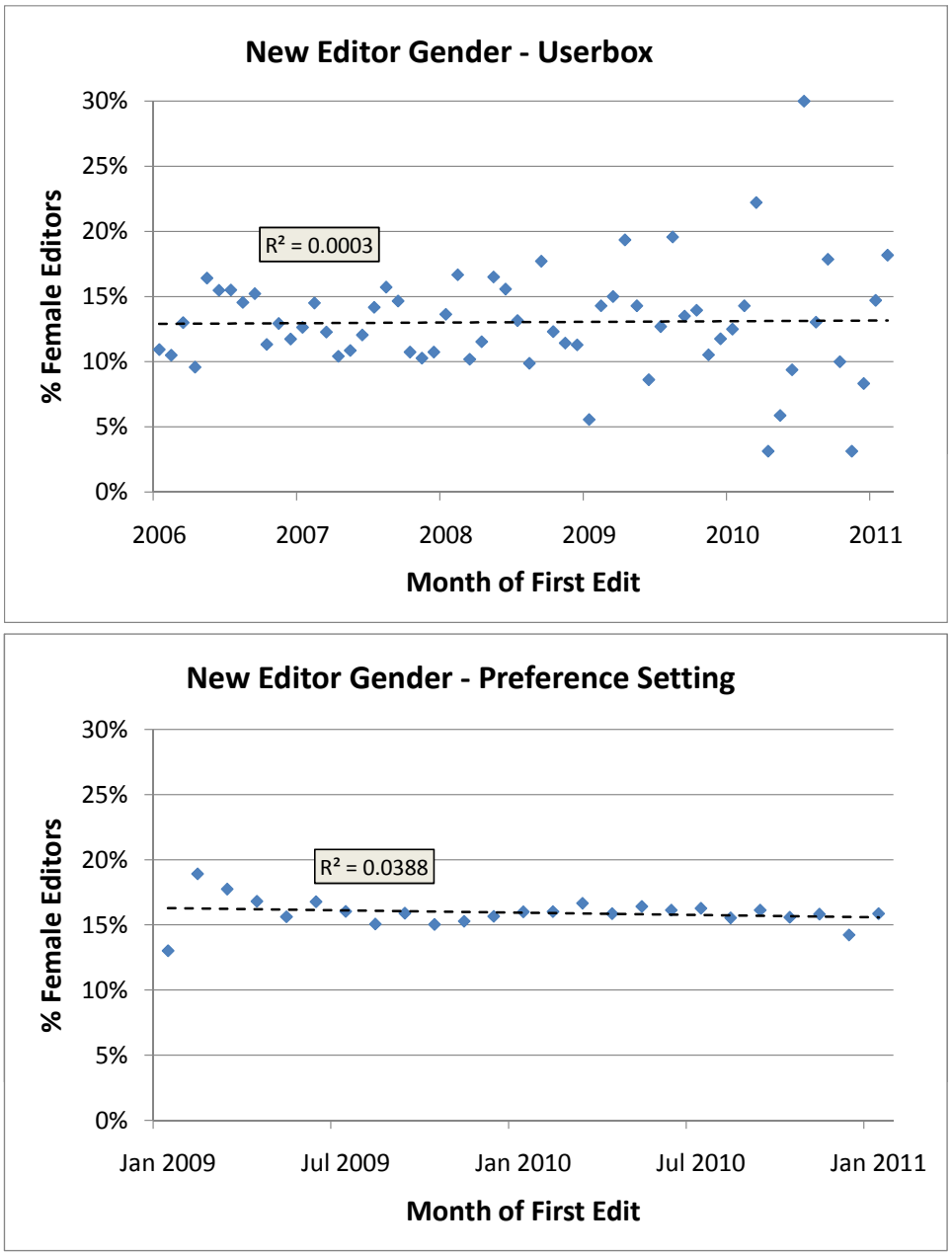

Figure 6.5: Wikipedia's gender gap in new editors as a function of time for users specifying their gender via userbox (top) and users specifying their gender via preference setting (bottom). Best-fit lines are plotted as black dotted lines, and have slopes of 0.0043% (top) and -0.029% (bottom) per month.

online gender gaps have been shrinking over time.

Note that the two charts indicate different female editor percentages (about 13% for userbox disclosures, and 16% for preference setting disclosures). This may be because the userbox gender disclosure method requires more Wikipedia knowledge to use. Specifically, one must (a) know about Wikipedia's user pages, (b) know about the gender userboxes, and (c) know how to place a gender userbox on one's user page. Thus, the survival differences shown earlier may lead to a smaller proportion of females who could specify their gender via userbox (that is, females might be more likely than males to stop editing Wikipedia before learning enough to use the gender userboxes).

Our findings support **H1a *Gap-Exists***, but not **H1b *Gap-Shrinking***. Wikipedia suffers from a substantial gender gap, but surprisingly, the gender gap has *not* been closing over time. Next, we turn to an exploration of how the gender gap is affecting Wikipedia.

6.4.2 RQ2: Gap-Matters

H2a *F-Coverage-Worse*: Coverage of topics with particular interest to females is inferior to topics with particular interest to males.

To examine the effect of the gender gap on content coverage quality, we performed two analyses: a general one based on editor activity, and a domain-specific one using an external data source that allowed us to more carefully control for possible confounds. For measuring coverage quality, we used article length as a proxy. While article length is a very simple metric, existing research shows that it is an excellent predictor of a Wikipedia article's assessment level [14, 171], which is a community-assigned rating of an article's overall quality.

Wikipedia-wide comparison. In our first analysis, we used the gender of an article's editors to determine whether the article topic is of male or female interest. Specifically, we defined an article's f_e as the proportion of its known-gender editors that are female. As we found earlier in section 6.4.1, edits by female editors turn out to be more rare than one would expect, so this metric is subject to high relative variance and noise. To help reduce the effect of noise, we limited this analysis to high-activity articles where we knew the gender of at least 30 editors. In addition, we excluded any articles that are less than 100 bytes long because such "articles" are likely to be redirects that point to other articles. Applying these constraints left us with 59,579 articles in the *Main* namespace, which is where all of Wikipedia's encyclopedic content

is located.

We considered an article topic to be "male" if it is in the bottom quintile of f_e, "neutral" if it is in the third (center) quintile, and "female" if it is in the top quintile. We found that the average male article is 33,301 bytes long, the average female article is 28,434 bytes long, and the average neutral article is 36,511 bytes long. All pair-wise differences in article length among these three quintiles are statistically significant (t-Test, $p < 0.001$). So, on average, male articles are significantly longer than female articles, which suggests that coverage quality of topics with particular interest to females is indeed lacking.

The results also indicate that neutral articles are of higher quality than both male and female articles. This is perhaps because gender-neutral topics appeal to both genders, and thus, are likely to garner the most overall attention from Wikipedia's editors.

Movie article comparison. Our second analysis for this hypothesis looked more deeply at one specific topic domain: movies. We chose to focus on this particular domain because it allowed us to use a methodology that let us more carefully control for other factors that may affect an article's length such as the popularity or age of the topic. Furthermore, this methodology is *not* subject to the Wikipedia data assumptions described in section 6.3.2.

For this analysis, we used data from the movie recommender web site *MovieLens*[3]. MovieLens users can assign ratings to movies in order to receive personalized movie recommendations. To date, MovieLens has collected over 15 million movie ratings from its 150,000 users. A key feature of this data set is that it contains self-reported gender information from over 80% of users who started using MovieLens before May 2003 (MovieLens stopped requesting demographic information from new users in May 2003). While MovieLens also appears to be affected by a gender gap (32% of its users are female), it is less imbalanced than Wikipedia, which allowed us to compute a gender metric even for relatively obscure movies. We defined a movie's f_r as the proportion of its known-gender raters that are female. To help avoid confounds due to gender differences in long-term MovieLens usage, we limited our analysis to movies that existed in the system as of May 2003.

We mapped each MovieLens movie to its corresponding article by scanning Wikipedia for articles that have a link to the movie's *IMDb* page and then applying basic heuristics to compare the movie name and article titles. We hand-checked 100 randomly-selected mappings and found no errors. We excluded any movies with fewer than ten known-gender raters, as well as movies

[3]http://movielens.umn.edu

that we could not locate a Wikipedia article for (3.7% of movies). The resulting movie data set contained 5,850 movies.

For this analysis, we built a linear regression model that predicts article length from f_r and several movie properties that may affect article length. The regression variables are summarized below. All variables except *Movie Age* are standardized with a z-score transformation. VIF values for these variables are below 1.3, so multicollinearity is not an issue.

- ***Article Length*** is our dependent variable, and is defined as the length (in characters) of the Wikipedia article about the movie (log-transformed for normality).
- ***Movie Gender*** is our independent variable, and is defined as f_r. We additionally include its quadratic term (labelled as "*Movie Gender Sq.*") because our previous analysis suggests that article length may have a non-linear relationship with topic gender.
- ***Movie Popularity*** is a control variable defined as the number of MovieLens ratings that the movie has (log-transformed for normality). Articles about well-known and often-rated movies may draw more attention from editors, and thus, may be longer.
- ***Movie Quality*** is a control variable defined as the average rating assigned to the movie by MovieLens users. Articles about highly-rated movies may be longer than those about poorly-rated movies, again due to increased editor attention.
- ***Movie Age*** is a control variable defined as the number of years since the movie's release date. Article length may vary with movie age due to better availability of information about newer movies.

The results of the regression model are shown in table 6.1. We see that even when controlling for variables that we expect to affect article length, both f_r and its quadratic term are significantly associated with article length. A plot of the effect size of movie gender is shown in figure 6.6. All else being equal, articles about "female" movies are shorter than ones about "male" movies. [4] We also built a similar regression model using WikiProject Film's article assessment ratings[5] as the dependent variable (coded as an equally spaced ordinal variable) and obtained qualitatively similar results; that is, "female" movie articles have lower assessment ratings than "male" movie articles.

[4]The Movie Popularity control variable has the expected effect. The Movie Quality variable has no significant effect. The Movie Age variable has the opposite of the expected effect: older movies tend to have longer articles. Perhaps only very noteworthy older movies were in our data set.

[5]http://en.wikipedia.org/wiki/WP:FILMA

Variable	Coef.	SE
Intercept	-0.335 ***	0.0190
Movie Popularity	0.676 ***	0.0106
Movie Quality	0.00521	0.0107
Movie Age	0.0123 ***	0.000604
Movie Gender	-0.144 ***	0.0107
Movie Gender Sq.	0.0257 ***	0.00692
Adj. $R^2 = 0.4704$, $F(5, 5843) = 1040$, $p < 0.001$		

Table 6.1: Results of the multiple linear regression model with movie article length as the dependent variable. Positive coefficients indicate variables associated with increased article length, and negative coefficients indicate variables associated with decreased article length (*** $p < 0.001$).

The results of these two analyses show the same thing: there are measurable gender-associated imbalances in Wikipedia's content coverage quality. The dearth of female editors appears to have led to female-interest topics receiving less attention and thus, lower-quality coverage in the encyclopedia.

In [127], Reagle and Rhue find a similar phenomenon: compared to *Encyclopedia Britannica*, biographies of women are more likely to be omitted or missing from Wikipedia than biographies of men. So, not only does Wikipedia have worse coverage of female-interest topics, it also has less consistent coverage of female subjects.

H2b *F-Social*: Females are more likely to be involved in social- and community-oriented areas of Wikipedia.

To test this hypothesis, we performed comparisons of male and female involvement in two areas of Wikipedia that represent increased social or community engagement. We describe each of these in turn below.

Editing Behavior. First, we looked at social engagement at a broad level by examining activity within the *User* and *User Talk* namespaces. Wikipedia's guideline[6] on the use of these namespaces states: "There is no fixed use for user pages, except that usually one's user page has something about oneself, and one's talk page is used for messaging." Thus, since pages in these namespaces are intended for self-expression and interpersonal communication, we interpret editing activity within them to be indicative of social engagement.

[6]http://en.wikipedia.org/wiki/WP:UPYES

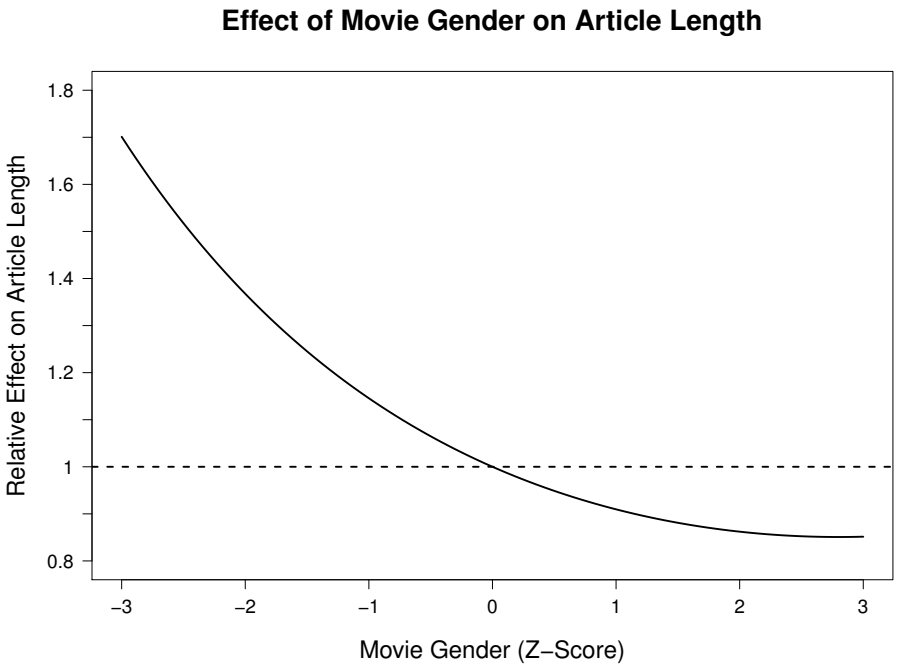

Figure 6.6: Effect size of movie gender (f_r) on Wikipedia article length. The x-axis shows movie gender as a standardized z-score, where negative values indicate male-skewed audiences, and positive values indicate female-skewed audiences. The y-axis shows the effect on article length as a multiplicative factor. For example, a movie that is at -2 on the x axis is two standard deviations more "male" than the average movie, and its Wikipedia article is expected to be 1.4 times as long as the average movie's article.

We looked at users with at least ten edits (4,990 females and 43,850 males), computed the percentage of each user's edits that are in each namespace, and compared the male and female means. Table 6.2 shows our findings for three groups of namespaces: 1) *User* and *User Talk*, 2) *Main* and *Talk*[7], and 3) all other namespaces[8]. We see that on average, a female makes a significantly higher concentration of her edits in the *User* and *User Talk* namespaces, mostly at the cost of fewer edits in *Main* and *Talk*.

Administrators. Second, we looked at how often editors have become Wikipedia administrators. Being an administrator provides a user with additional capabilities such as protecting pages, hiding revisions from public view, and blocking others from editing. Besides their usual

[7]These namespaces contain encyclopedic content and discussions about the content, respectively.

[8]These include supporting materials such as site policies, multimedia files, and help pages.

Namespace	Females	Males
User & User Talk ***	**25.2%**	19.1%
Main & Talk ***	69.1%	**74.5%**
Other namespaces ***	1.88%	**2.82%**

Table 6.2: Comparisons of Wikipedia editing behavior across different namespaces. The figures are the percentage of edits occurring in each namespace, averaged across female or male editors. Differences were compared using a t-Test (*** $p < 0.001$).

	Females	Males
All Wikipedia Editors ***	0.33% (50/15,362)	**0.59%** (579/98,486)
Users with 2,000+ edits *	**18.6%** (49/263)	13.4% (575/4,283)

Table 6.3: Comparison of gender and Wikipedia administratorship. The figures are the percentage of each population who are administrators. Proportions were compared using a Chi-square test (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$).

editing tasks, administrators have the additional responsibility of being Wikipedia's janitors and custodians[9]. Administrators help clean up after vandals, resolve conflicts between editors, and maintain order in the community. Since becoming an administrator is a major form of civic duty in Wikipedia, we consider it an indicator of increased community engagement.

Looking at table 6.3, we see that overall, a greater proportion of males than females are administrators. However, one major factor in successfully becoming an administrator is having a substantial edit count [19], and we know from our earlier results that females are more likely to leave Wikipedia before accumulating many edits. If we restrict our analysis to users who have at least 2,000 edits (all but five administrators have over 2,000 edits), the tables are turned. Within this group of dedicated Wikipedians, females are actually significantly *more* likely to be administrators than their male counterparts. We also performed the same comparison using thresholds of 1,000 and 4,000 edits and obtained the same qualitative result.

The data support both hypotheses under **RQ2: Gap-Matters**. In **H2a *F-Coverage-Worse***, we find that the gender gap appears to have a detrimental effect on content coverage of topics with particular interest to females. Our results for **H2b *F-Social*** suggest that addressing the gender gap could help Wikipedia better address its needs in social- and community-oriented areas.

[9]Wikipedia's symbol for an administrator is a janitorial mop!

6.4.3 RQ3: Gender-Conflict

H3a *F-Uncontentious*: Females tend to avoid controversial or contentious articles.

We address this hypothesis by looking at the edit protection status of *Main* namespace articles that have a high concentration of either female editors or male editors. An edit-protected article cannot be edited by certain classes of users (typically new or anonymous editors, but several levels of protection exist). Wikipedia's protection policy[10] states that articles that are subject to content disputes, vandalism, or other forms of disruption are candidates for protection. Therefore, we consider protected articles to be representative of articles that are about controversial or contentious topics.

We found that 5.20% of the "female" articles described in section 6.4.2 are protected, while just 2.39% of the "male" articles are protected, $\chi^2(1, N = 23989) = 129.1$, $p < 0.001$. Thus, articles that have a higher concentration of female editorship are actually *more* likely to be contentious than those with more males.

H3b *F-Reverted-More*: Female editors are more likely to have their early edits reverted.

Recall that a revert occurs when a Wikipedia editor undoes another editor's edit, thereby discarding their contribution. Our analysis of revert behavior uses Wikipedia's January 2010 data dump, which is the most recent dump for which we had access to processed revert information. We used the method described in [122] to detect reverts and to classify whether they are for damage repair (specifically, we use the "D-Loose" classification, which is an imperfect heuristic, but is capable of identifying many common vandalism repair patterns). Because we were only interested in reverts of good-faith attempts to improve the encyclopedia, we only considered reverts in the *Main* namespace that were *not* for the purposes of repairing damage or vandalism. To limit the effect of right truncation, we also only considered reverts that occurred within one week of an edit (this was the case for over 95% of the reverts in our data set).

Using this data, we took each user's chronological sequence of *Main* namespace edits, partitioned the edits into bins of increasing size to represent different stages of editor tenure, and determined what percentage of edits in each bin were reverted for non-vandalism reasons. Table 6.4 shows the results of this analysis aggregated by gender. We see that in the first three bins, which consist of users' first seven edits, the average reverted edit percentage for females is significantly higher than that for males. Therefore, females are indeed significantly more likely

[10] http://en.wikipedia.org/wiki/WP:PROTECT

Revert Rates During Editor Tenure				
Edit #	**Females**		**Males**	
	N	**Mean**	**N**	**Mean**
1	6,305	6.99% ***	53,738	4.96%
2-3	4,989	6.04% ***	45,514	4.44%
4-7	3,798	4.69% **	37,272	3.98%
8-15	2,871	3.47%	30,066	3.12%
16-31	2,080	2.56%	23,798	2.66%
32-63	1,490	2.33%	18,270	2.49%
64-127	1,039	1.97%	13,850	2.27%
128-255	749	2.29%	10,355	2.07%

Table 6.4: Average rate at which editors are reverted for non-vandalism-related reasons, by gender and by stage of editor tenure (the first row shows the reverted rates for users' first edits, the second row shows the rates for users' second and third edits, and so on). Averages were compared using a t-Test (*** $p < 0.001$, ** $p < 0.01$).

than males to have their edits reverted during the early parts of their tenure.

Interestingly, beyond this initial handful of edits, we see little statistical difference between females and males in how often they are reverted. This suggests that females and males who manage to reach a modest level of Wikipedia experience are on par with each other with respect to community acceptance.

H3c *F-Reverted-Leave*: Female editors are more likely to stop editing and leave Wikipedia when being reverted as newcomers.

Finally, we look at how editors react to being reverted. The survival analysis in section 6.4.1 indicated that females appear to stop editing Wikipedia sooner than males. We investigate this phenomenon more deeply here, looking specifically at whether female newcomers are more likely than their male counterparts to stop editing if they are reverted. To do so, we developed a Cox regression model [31] to determine which factors are associated with longer (or shorter) activity lifespan. The variables in our model are summarized below. VIF values for these variables are below 1.4.

- ***Gender*** is the editor's gender, dummy-coded with females as 1 and males as 0.
- ***Edits24H*** is the number of edits made in the first 24 hours of editing Wikipedia (this is log-transformed for normality).

Variable	Coef.	Hazard Ratio	95% CI
Gender (female)	0.248	1.281 ***	1.229-1.335
Edits24H	-0.164	0.849 ***	0.838-0.861
%RvVandal	0.486	1.626 ***	1.498-1.766
%RvNonVandal	0.332	1.394 ***	1.314-1.478
%RvNV × Gen	0.0393	1.040	0.904-1.197
			Adj. $R^2 = 0.043$, $p < 0.001$

Table 6.5: Results of Cox proportional hazards regression model predicting activity lifespan for editors who started editing during 2009. Variables with hazard ratios above 1 are associated with shorter activity lifespans, while those with ratios below 1 are associated with longer lifespans (*** $p < 0.001$).

- ***%RvVandal*** is the proportion of edits made in the first 24 hours that were reverted for vandalism-related reasons.
- ***%RvNonVandal*** is the proportion of edits made in the first 24 hours that were reverted, but not for vandalism-related reasons.
- ***%RvNV × Gen*** is an interaction term between *%RvNonVandal* and *Gender*, and is used to study the interaction effect between gender and being reverted for non-vandalism reasons.

The results of the regression model are shown in table 6.5. The model has limited predictive power, but nonetheless, we see that all the variables *except* the interaction term have a significant association with activity lifespan. Making more edits during one's first 24 hours as a Wikipedia editor is associated with a longer activity lifespan, while having one's early edits reverted for any reason, vandalism-related or otherwise, is associated with a shorter lifespan. Note that even after taking these factors into account, being female *still* has a strong association with shorter activity lifespan.

However, contrary to our expectation, there is no interaction effect between gender and being reverted for non-vandalism reasons. It appears that males and females are affected similarly when their edits are not accepted by the Wikipedia community. This point deserves elucidation. Although reverts appear to drive newcomers away from Wikipedia, and although females are more likely to be reverted as newcomers, if a revert does happen to a female, the likelihood of her departure is not affected more than that of a male in a similar situation. Therefore, the gender gap appears to be due more to females being reverted disproportionately often, rather than to females reacting more strongly when they are reverted.

Hypothesis	Result	Description
H1a *Gap-Exists*	Supported	Wikipedia has relatively few female editors, and they leave Wikipedia sooner than males
H1b *Gap-Shrinking*	Unsupported	The gender gap has *not* been shrinking over time
H2a *F-Coverage-Worse*	Supported	Coverage of "female" topics is inferior to coverage of "male" topics
H2b *F-Social*	Supported	Females are more likely to participate in social- or community-oriented areas of Wikipedia
H3a *F-Uncontentious*	*Reversed*	Articles with high female editor concentrations are *more* contentious
H3b *F-Reverted-More*	Supported	Female newcomers are reverted more than males
H3c *F-Reverted-Leave*	Unsupported	Being reverted as newcomers has the same apparent effect on males and females

Table 6.6: A summary of our hypotheses and findings.

Our findings for **RQ3: Gender-Conflict** are mixed. In **H3b *F-Reverted-More*** we find that female newcomers have a more difficult time getting good-faith contributions to be accepted by the community. However, our findings for **H3c *F-Reverted-Leave*** indicate that the effect of having an edit reverted is no worse for females than it is for males. Unexpectedly, we find that female editors are *more* concentrated in areas with high controversy (**H3a *F-Uncontentious***). In summary, the available data indicate that female editors experience more adversity than male editors in all the areas that we studied.

6.5 Discussion

Table 6.6 shows a summary of our hypotheses, whether we found support for each hypothesis, and a brief statement of our results. One way of interpreting our results is in terms of the Reader-to-Leader Framework [121]. This framework, developed by Preece and Shneiderman, describes a process in which users of a social media system move through four levels of participation: reader, contributor, collaborator, and leader. We systematically look at female participation in Wikipedia through the lens of this framework.

Readers. Becoming a consumer of social media is a typical first step toward active participation. Existing survey research yields mixed findings about the share of females in Wikipedia's readership [96, 177, 48]. The most accurate and unbiased of these appears to be from Pew Research Center [177], which used random telephone dialing to draw a random sample, along

with statistical correction techniques to account for non-response bias. It indicates a female readership share of approximately 47%, which suggests that Wikipedia is effective at drawing a reasonably gender-balanced population of readers. The present research does not consider readership directly, but this figure will serve as a useful reference point.

Contributors. Perhaps the most challenging task for a social media system is to convince a reader to start giving back to the community – that is, to turn readers into contributors, or in Wikipedia's case, editors. Our results for **H1a *Gap-Exists*** show that Wikipedia is much less successful in "converting" female readers than male readers, dropping from a 47% female share in readership to a 16% share in editors, a figure that has shown little to no change for years according to our results for **H1b *Gap-Shrinking***.

Collaborators. When multiple contributors come together to work toward a common goal, they become collaborators. In researching **H2b *F-Social*** we found that females edit more in the *User* and *User Talk* namespaces, indicating a potential interest in collaboration. Our data do not contain metrics that directly describe collaboration activity, but a reasonable proxy is simply the presence of sustained editing activity. Our survival analyses in **H1a *Gap-Exists*** and **H3b *F-Reverted-More*** indicate that females who become contributors stop editing Wikipedia sooner than males. Furthermore, both **H3a *F-Uncontentious*** and **H3b *F-Reverted-More*** suggest that females encounter more adversity in Wikipedia. Together, these data suggest that while females appear interested in becoming collaborators, they have more difficulty in making the transition for a variety of reasons.

Leaders. Effective collaborators who are passionate about their work and who are interested in the system at a meta-level emerge as community leaders. Preece and Shneiderman specifically examine Wikipedia administration as an exemplar of a leadership role in online communities. In our analysis for **H2b *F-Social***, we see that females who reach a high level of participation are more likely than their male counterparts to take on a leadership and administration role. However, as we saw in **H1a *Gap-Exists*** only 6% of the editors who have contributed more than 2,000 edits are female. [11] Some Wikipedians have observed that an administrator shortage may be looming as current administrators "retire," but few new administrators are emerging to fill their shoes. [12] Addressing the gender gap in high-participation editors might be an opportunity to meet this demand for more administrators.

[11] Nearly all administrators have at least 2,000 edits.

[12] http://en.wikipedia.org/w/index.php?oldid=393297323

Implications. Overall, our findings indicate that there is a substantial male-skewed gender imbalance in English Wikipedia editors that does not appear to be closing at any appreciable rate. This is at odds with observed participation rates in other forms of online social media that are gender-balanced or that are even female-skewed [93, 117]. Furthermore, we find that the gender gap matters to Wikipedia: there are gender-associated imbalances in coverage quality, which impinges on Wikipedia's goal of producing a high-quality encyclopedia. Not only would addressing the gender gap help resolve the quality disparity, it would also help increase diversity within Wikipedia's collaborations, which may improve group productivity and retention rates [22] as well as decision-making quality, as shown in chapter 5.

The problem is subtle, and simple attempts at solution without detailed understanding are likely to fail. Of our seven hypotheses, all of which seemed plausible before we began our study, three were not supported by the data, sometimes in quite surprising ways. How can it be that the gender gap in Wikipedia is not closing, though overall Internet usage has become gender-balanced?

Taken together, our results in ***RQ3: Gender-Conflict*** hint at a culture that may be resistant to female participation. More research, including interviews, surveys, and focus groups is needed to determine the underlying causes of the problems evidenced by our findings, and to determine what can be done to improve the situation. We hope this research is a first step toward addressing the gender imbalance – and the problems it causes – in Wikipedia.

Chapter 7

Shilling: Deviance in Social Production Communities

7.1 Introduction

Harnessing the altruism and social capital of members of SPCs can lead to the creation of vast and useful resources, as we have seen with Wikipedia. However, relying on goodwill in this way can leave the door open to less scrupulous individuals who may try to manipulate the community for individual gain. Throughout the work in this thesis, we have often touched upon ways that SPCs are vulnerable to self-serving forms of manipulation.

In chapter 3, we saw that many of Wikipedia's new articles involve inappropriate content, including cases where editors use Wikipedia as a marketing platform by creating encyclopedia articles that are commercial advertisements. When we looked at collaborative curation in MovieLens in chapter 4, we expected to find similarly deviant and self-serving behavior when we provided users with the ability to directly curate the movie database (though our results indicated no such behavior). Our look at small-group decision making in chapter 5 revealed possible evidence of strategic behavior when Wikipedia editors formed decision-making groups.

We close our exploration of challenges in collaborative curation with a deeper look at one specific way that malicious users can manipulate the technical mechanisms in an SPC: submitting false opinions and reviews to a recommender system in order to manipulate what things other users see and the choices that they make.

One issue that many users face in social production communities is that of *information overload*. Sourceforge has hundreds of thousands of projects, Wikipedia has millions of articles, and Flickr has billions of photos. Users are overwhelmed by the amount of information available, and cannot possibly evaluate all the possible choices when looking for something interesting to consume.

In recent years, recommender systems have emerged as one tool that can help people overcome this problem and quickly locate interesting items. These systems are based, in part, on principles of collaborative curation. They collect judgements and opinions about items from the community, and aggregate them in order to make recommendations to a user regarding which items she may find interesting. One instance of a recommender system is *MovieLens*, which we used as a research platform in chapter 4.

While recommender systems offer clear benefits to users, they can also be a valuable asset to retail companies in helping their customers find products that they might want to buy and, in effect, increasing not only sales, but perhaps also cross-sales and customer retention. This is particularly true in the realm of e-commerce. For example, *Amazon.com* has made many recommender systems available to their customers. These range from manually-operated recommenders where users write reviews or create lists of recommended products, to automated systems that recommend items based on observed browsing or purchasing patterns.

A producer (e.g., author, composer, or inventor) naturally would like his or her own products to do well in the marketplace. In the context of a recommender system, there is, therefore, a motivation to want one's own products to be recommended more often than competitors' products. Of course, one way to accomplish this is to produce quality goods that people like and regard highly. However, unscrupulous producers may opt to take a more deceitful route; they may try to influence recommender systems in such a way that their items are recommended to users more often, whether or not they are of high quality.

An instance of a company generating false "recommendations" to consumers arose in June 2001 when Sony Pictures admitted that it had used fake quotes from non-existent movie critics to promote a number of newly released films.[1] The online retailer *Amazon.com* has found that their recommenders are prone to some level of abuse on at least two different occasions.[2,3] Also, *eBay*, which uses a recommender system as a reputation mechanism to help users know which

[1] http://news.bbc.co.uk/1/hi/entertainment/film/1368666.stm
[2] http://www.wired.com/news/ebiz/0,1272,53634,00.html
[3] http://news.com.com/2100-1023-976435.html

sellers they can trust, has found itself contending with users who subvert their system in various ways, including people who purchase good ratings (feedback) from other members in order to bolster their own reputations.[4]

One way to influence a recommender system is to arrange to have a group of users enter the system and vouch for the items in question. These users become *shills*, whose false opinions are intended to mislead other users. If successful, shills pose a serious threat to users and operators of recommender systems. They may cost users time and money by recommending bad items. They may cost operators by degrading the user's level of trust in the recommender system and the company running it.

This chapter focuses on shill attacks on recommender systems that use automated collaborative filtering (ACF) to generate recommendations. ACF is a class of algorithms commonly used in the implementation of recommender systems. These algorithms operate on the basis that similar users have similar tastes; thus, if people similar to you can be located, then the items they enjoy are likely to be ones you will also enjoy. These algorithms normally have two modes of operation: prediction and recommendation. In the *prediction* mode, the algorithm simply predicts how much a user will like some item or set of items. The items may have been selected by the user through browsing or searching. In the *recommendation* mode, the algorithm produces an ordered list of items that it believes the user is most likely to enjoy. This distinction will become important as we explore the practical effects of shilling attacks on recommender systems.

7.1.1 Prior Work

Recommender Systems

One of the earliest instances of a collaborative filtering system is *Tapestry* [49], an experimental email system that allowed its users to apply collaborative filtering to decide which documents to read. Resnick et al. introduced an ACF algorithm based on k-Nearest-Neighbor and applied it to messages in Usenet newsgroups [129]. Since then, a number of improvements to this kNN algorithm have been proposed [67, 54]. The *user-user* algorithm that we study is a tuned version of Resnick et al.'s kNN algorithm. A related ACF algorithm that we study is a tuned

[4] http://pages.ebay.com/help/policies/feedback-manipulation.html

version of the item-based k-Nearest-Neighbor algorithm described in [133]. We call this algorithm the *item-item* algorithm in this work. Many other ACF recommendation algorithms have been developed as well, including ones based on singular value decomposition [134], Bayesian networks [15], and factor analysis [21].

Besides *MovieLens*, many other recommender systems exist, particularly on e-commerce sites. Schafer [135] examines and categorizes a large set of these commercialized recommender systems. In addition, numerous recommenders in a variety of domains have been developed for research purposes, including *GroupLens* (Usenet news) [105], *Ringo* (music) [139], and *Jester* (jokes) [50].

Attacks on Recommenders

While a large body of work exists on recommender system algorithms, less attention has been devoted to exploring and improving their resistance to attacks. Dellarocas [32] outlines several attacks on reputation systems used in online trading communities such as *eBay* and proposes a predictive algorithm similar to existing collaborative filtering algorithms which helps minimize the effect of the attacks. Canny [21] presents a system in which user preferences (ratings) are kept private both from the recommender system's administrators and from other users, which he believes can mitigate attacks where the system administrators are involved (e.g. a retailer being paid to place a company's items highly on recommendation lists).

O'Mahony et al. [112] performed empirical studies of the resistance of the kNN user-user algorithm to attacks based on injecting a number of shill users into the system. The attacks were shown to be successful both at *push*-ing items by raising predicted ratings, and at *nuke*-ing items by lowering predicted ratings. Furthermore, they present a theoretical analysis of the effect of noise – perhaps injected by shills – on the performance of ACF algorithms and perform several experiments with real-world data sets to evaluate the generated models. This work builds on existing work by including more recommender algorithms and by evaluating the attacks on recommendations as well as on predictions.

Impact of Successful Attacks

Psychologists have shown experimentally that people tend to conform with the behavior of their peers, even when those behaviors may be wrong. For instance, consider the classic conformity

study by Asch [6] where test subjects were asked to answer simple questions in small groups. In the experimental condition, groups consisted of experimental confederates who deliberately gave incorrect responses, and even though the correct answer was obvious, test subjects agreed with the group and provided an incorrect response in one-third of all trials.

A similar effect is likely to occur in interactions between computers and people. Nass and Moon [110] found that people tend to treat computers as they would treat people – in one of their experiments, they showed that people are less likely to criticize a computer's performance if that computer asks for the evaluation than if another computer asks.

Cosley et al. [30] found that users are indeed affected by manipulated predictions provided by a recommender. Users' ratings were affected by the presence of a predicted rating display, even when that prediction was manipulated. When an artificially inflated prediction was displayed, users provided higher ratings, and likewise, they provided lower ratings when the prediction was manipulated downward.

Whether this effect represents a genuine change in opinion is unknown – it might just be that users conform in what they say, not in what they actually believe. Nonetheless, Cosley et al.'s results suggest that shilling may be doubly dangerous: if shilling attacks can lead users to submit higher ratings for some item, then there may be a cascade effect where the ACF algorithm produces even more high recommendations for that item to other users.

7.1.2 Hypotheses

The previous research on shilling demonstrates that shilling should be of concern in recommender systems, but leaves several important questions unanswered. The first of these questions is about how effective shilling is on finely tuned versions of the user-user [67] and item-item [133] algorithms. Though the algorithms are based on fundamentally similar ideas about relationships between users and items, their implementations are sufficiently different that we suspect they may exhibit different behavior under shilling attack. Formally:

Hypothesis 1 *Different ACF algorithms respond differently to shilling attacks.*

This hypothesis has importance for designers of ACF algorithms, who may be able to design algorithms that are more resistant to attack, for operators of recommender systems, who may be able to select algorithms that are resistant to likely attacks, and for evaluators of shilling attacks, who may need to be aware their results are algorithm dependent. We suspect there

will be differences in shilling resistance among other less similar recommender algorithms as well; if we find differences between these two algorithms we will recommend future work to understand in detail the shilling resistance of the entire suite of known recommender algorithms.

Recommender systems in e-commerce, where shilling is most often feared [32, 112], are used more often to produce recommendations [135] than predictions. Attacks should be judged based on how they affect the recommender in its most common mode of operation – after all, if a user only looks at the top several items on a list, does it matter that the shilling attack has changed what she would have seen for the 500th item on her list?

Hypothesis 2 *Shilling attacks affect recommender algorithms differently from prediction algorithms.*

If this hypothesis is supported, it will mean that researchers who evaluate attacks must base their measure of the effectiveness of the attack on the specific ways the targeted recommender system is being used in practice. Metrics for evaluating recommender algorithms have traditionally been focused on predictions; perhaps for recommendation tasks recommendation metrics should be used instead.

The fact that past research has shown that shilling is effective in some cases, raises the question of whether the operators of recommender systems can detect that their systems are under effective shilling attack. Our next hypothesis is that they cannot do so with existing tools. Note that this hypothesis is distinct from hypothesis 2 even though both are about metrics. Hypothesis 2 is about measures that evaluate shilling attacks based on knowing exactly which items are being attacked, which the operator of the recommender system will not know.

Hypothesis 3 *Shilling attacks are not detectable using traditional measures of algorithm performance.*

ACF algorithm designers often utilize metrics such as Mean Absolute Error (MAE) to evaluate the overall predictive accuracy of their algorithms and to compare it with other algorithms. Thus, it may be tempting to use such metrics to detect attacks by looking for changes in algorithm quality caused by attacks. We believe this will generally not be possible; that is, attacks can be subtle and focused enough that their overall effect on the system is minimal.

This hypothesis is unnerving to those who – like us – run a recommender system, since it means that our systems may already be under successful attack, and that despite all of the

measurement tools at our disposal we may be unaware of all but the crudest attacks. We seek to spur understanding of which evaluation techniques are best for detecting shilling attacks.

One possible place to look for detecting shilling attacks is to understand which target items are most vulnerable to attack. We hypothesize that the size and shape of an item's ratings distribution can influence its vulnerability to shilling attack.

Hypothesis 4 *Ratings distribution of the target item influences attack effectiveness.*

We will study popularity (number of ratings), likability (average rating), and entropy (a measure of rater agreement) as the variables that describe the ratings distribution. We believe that the following statements will be true:

- Popularity – the less popular an item is, the easier it is to manipulate the predictions and recommendations for that item
- Likability – the more well-liked an item is, the easier it is to cause that item to be recommended more often
- Entropy – the higher the entropy of an item's ratings, the easier it is to manipulate the predictions and recommendations for that item

7.1.3 Contributions

Our analyses of shilling attacks build on the work presented in [112] and further explore the feasibility and effectiveness of influencing recommender systems based on ACF algorithms.

First, we propose a set of dimensions that describe and categorize a wide variety of shilling attacks. These dimensions set the stage for the types of evaluation we carry out, though our evaluation so far is only of the set of dimensions we expect to be most important in practice.

Second, we build on past work to develop two basic attack types, and perform a series of experiments to study their effectiveness in influencing both the predictions and the recommendations made by two different ACF algorithms. In our experimental work we simultaneously evaluate all three aspects of the shilling attack: the attack itself, the ACF algorithm under attack, and the different metrics used to evaluate both the algorithms and the attacks.

Finally, we present the results of the experiments and examine how they support or refute our hypotheses, and conclude with a look at some open questions and possible future work.

7.2 Dimensions of Attacks

In this work we consider *shill attacks* where the attacker's only available action is to introduce a new set of users and a set of ratings made by those new users to the recommender system. In the case of online recommenders, traditional attacks such as denial-of-service, password cracking, system hacking, and bribery are possible, but are beyond the scope of our analysis.

Each shill attack has a number of intrinsic properties that can be useful in describing and comparing different attacks. A list of these properties, or dimensions, follows.

7.2.1 Attack Intent

Different shill attacks may have very different intents. While the direct result of a shill attack is generally that the predictions made to users are manipulated in some way, an attacker's goal can be one of several alternatives. Two straightforward intents are to "push" one or more items in the system in order to have them recommended to more users and, conversely, to "nuke" a set of items to cause them to be recommended to fewer users.

Another possible intent is to simply damage the recommender system as a whole; that is, to reduce prediction and recommendation quality across the board with the goal of causing users to stop trusting the system and eventually to stop using it. A successful attack of this nature might benefit competing recommender systems.

7.2.2 Targets

Shill attacks can be directed at a particular subset of users and a subset of items in a recommender system. It is in an attacker's best interest to restrict the effect of an attack to a small target set of items in order to be more subtle and try to avoid detection by the system operators. Additionally, it might be beneficial to also restrict the effect to some desirable set of target users. For instance, it could induce suspicion to cause a rap album to be recommended to a connoisseur of classical music who would have no plausible interest in such an album.

7.2.3 Required Knowledge

Attacks may require some level of knowledge about the items, users, ratings, and algorithms in the recommender system being attacked. An informed attack will generally be more effective

than an uninformed attack. Further knowledge about the system such as ratings sparsity, ratings distribution, and ACF algorithm parameters can help in choosing which attack to employ and in tuning attack parameters to maximize effectiveness and minimize detectability.

7.2.4 Cost

A shill attack has an associated cost that depends on the level of effort and information needed to successfully execute the attack. With a cost dimension and a suitable means of evaluating attack effectiveness, one might be able to evaluate attacks on a cost/benefit basis to determine if it is economically worthwhile to execute an attack. The following factors contribute to the cost of a given attack:

- Size of attack: the number of new users and ratings.
- Difficulty of interacting with the recommender system. For instance, an attack on a system that employs anti-automation techniques such as CAPTCHAs may have a higher cost than an attack on a system that does not [163].
- Obtaining required knowledge about the algorithm, users, items, and ratings in the recommender system.
- Any other resources required for attack planning or execution, such as additional logistical, computational, or technical requirements.

7.2.5 Algorithm Dependence

Some shill attacks may be specifically designed to exploit a particular weakness in a specific algorithm or class of algorithms, while others might be more general and can be effective against a variety of algorithms. More specific attacks will likely require fewer resources for the same effectiveness, but require detailed knowledge of the algorithm being used and its parameter settings.

7.2.6 Detectability

Inherent properties of attacks may make them more or less easily detectable, both to users and operators of the recommender system. In general, an attacker would like to be less detectable to

operators in order to be able to sustain the attack for as long as possible before being discovered and stopped. The importance of detectability to users depends on the attack intent. This dimension can evolve very rapidly as shill detection methods are developed or improved, similar to the arms race seen in spam (unsolicited commercial email) detection and detection evasion.

7.3 Experimental Design

7.3.1 Shill Attack Design

This work focuses on algorithmic attacks that attempt to push or nuke an item (that is, raise or lower the recommender's predictions for the item) by introducing shill users into the system. It is assumed that the attacker does not have access to the ratings matrix, but can obtain broad statistical measures of the ratings data. We do not concern ourselves with reducing the cost of the attacks or with designing attacks that are difficult to detect. If these brute force attacks prove difficult for operators to detect, recommender systems operators should be concerned indeed!

We begin with the type of attacks used in [32]. These attacks inject a collection of new users into the system, each of which has rated a set of items to try to be similar to existing users, and has rated the particular item being attacked very high in order to push it. Practical implementations of both user-user and item-item algorithms scale correlations according to the number of ratings in common [67]. [112] observes that this type of attack does not work well for such implementations if the shill users rate too few items because their correlations with target users become too low after scaling. We therefore modify the attack to rate *all* movies in the system to maximize the number of items in common between shill users and real users.

Attacks like these are related to *filterbots*, which rate all items using some established heuristic [54]. However, filterbots are used to improve recommendation coverage and quality, while these shills ("shillerbots," perhaps) will be used to directly manipulate predictions for a small subset of items.

One way in which the ACF algorithms we test are different from those used in the past [32, 112] is that we do not use negative correlations between items. We made this decision because our past experience has been that negative correlations often lead to recommendations that are inconsistent with user preferences. Shills might exploit negative correlations to produce very strong attacks – but system operators will likely disable the use of negative correlations to improve quality first.

Our attacks target the entire population of users and a small target set of items while requiring relatively limited knowledge about the ratings matrix. The number of new shill users introduced to the recommender system is varied between 25 and 100. The intent of these attacks is to either push or nuke the target set of movies. The two attack methods developed are:

RandomBot

RandomBots are a naive attack in which each introduced user rates items not in the target set randomly on a normal distribution with mean 3.6 and standard deviation 1.1. These values are chosen because they represent the ratings distribution in the data set. Even if these values are not known to the attacker, they can be estimated relatively easily, perhaps by observing people using the recommender system and obtaining a sample of their ratings. A normal distribution is used to approximate the observed user rating behavior in *MovieLens*. To accomplish its objective, the filterbot rates items in the target set with value equal to either the minimum or maximum allowed rating, depending on its intentions (nuke or push, respectively).

AverageBot

AverageBots are a somewhat more sophisticated attack than RandomBots and require knowledge of the average rating of each item in the system. A number of recommender systems, including *MovieLens*, will readily provide this information. Furthermore, such aggregate information about users' preferences can commonly be found from other sources. In the movie domain, the *Internet Movie Database* (http://www.imdb.com) publicly displays the average user ratings of listed movies.

Each introduced user rates items not in the target set randomly on a normal distribution with mean equal to the average rating of the item being rated and standard deviation 1.1. The intuition behind this attack is that providing ratings centered around the average rating will help the filterbot be more similar to existing users, and thus, have a larger effect on the recommendations. As with the RandomBot, items in the target set are rated with a minimal or maximal value depending on the attack intent.

7.3.2 Data Set

A data set derived from *MovieLens* consisting of 999,799 ratings on 3,404 movies by 7,463 users is used in all of our experiments. All ratings are integral values between 1 and 5, inclusive, where 1 represents a poor movie and 5 represents an excellent one.

7.3.3 ACF Algorithms

We experimented with two commonly-used ACF algorithms – kNN user-user and kNN item-item.

kNN User-User

We chose the classic kNN user-user algorithm introduced by Resnick et al. in [129] because it is considered to be a good baseline algorithm and is widely used in both academia and industry.

The user-user algorithm defines the predicted rating $p_{u,i}$ for a user u on an item i as follows.

$$p_{u,i} = \overline{r_u} + \frac{\sum_{v \in U_{u,i}} [w_{u,v}(r_{v,i} - \overline{r_v})]}{\sum_{v \in U_{u,i}} |w_{u,v}|} \tag{7.1}$$

Here, $r_{u,i}$ is user u's rating on item i, $\overline{r_u}$ is user u's average rating over all rated items, $w_{u,v}$ is the mean-adjusted Pearson correlation ("similarity") between users u and v, and $U_{u,i}$ is user u's neighborhood with respect to item i and consists of the k users who have rated i and have the greatest Pearson correlation with u. k is a tunable parameter and represents the number of neighbors.

Several optimizations and suggested parameters from [67] are used. In particular, we set k to 20 and use the $n/50$ significance weighting and deviation from mean optimizations. A similarity threshold of 0.1 is also used. The only difference between the user-user variant used in this work and the published algorithm is that only positive similarities are considered here; negative similarities and ones under the threshold are ignored. This variant was also used for several years in the *MovieLens* recommender system and does not appear to be detrimental to performance.

kNN Item-Item

The kNN item-item algorithm was introduced by Sarwar et al. [133] and is similar to the user-user algorithm in both definition and predictive accuracy. However, it computes and uses similarities between items rather than users. We choose to experiment with this algorithm in order to explore how well an attack that is effective against user-user operates with a somewhat different algorithm. The item-item algorithm produces predictions as follows:

$$p_{u,i} = \frac{\sum_{j \in allsimilaritems}[s_{i,j} * r_{u,j}]}{\sum_{j \in allsimilaritems} |s_{i,j}|} \tag{7.2}$$

Here, $s_{i,j}$ is the similarity between items i and j. The algorithm implementation used in this work is the *MultiLens* recommender engine developed by Miller [106], which uses the adjusted cosine method of computing similarity, and considers the 20 rated items with highest similarity to be the set of "all similar items" (again, ignoring negative similarities). A tuned version of this algorithm is currently utilized in *MovieLens*.

7.3.4 Methods

A total of twenty-four experiments were performed in a 2x2x2x3 design. The algorithm (user-user or item-item), attack type (AverageBot or RandomBot), attack intent (nuke or push), and number of new users/bots (25, 50, or 100) were varied in each experiment.

The target set for the experiments consists of 22 items. This set was selected to include a variety of different movie types including future releases, new releases, obscure films, popular films, controversial films, and long-standing favorites. In terms of ratings properties, this selection of items represents a wide range of popularity, entropy, and likability. Table 7.1 displays the properties of items in the target set.

7.3.5 Metrics

There are two things that we would like to be able to measure:

1. How much the overall accuracy of the ACF algorithm is affected by an attack.

2. How effective an attack is in accomplishing its goal.

Table 7.1: Properties of movies in chosen target set. Ratings is total number of ratings (popularity), mean is average rating (likability), and entropy is the standard information-theoretic entropy of the ratings distribution. Recall that ratings are provided on a 5-point scale.

Item	Ratings	Mean	Entropy
1	0	N/A	N/A
2	7	2.57	0.99
3	8	2.88	1.56
4	17	3.00	2.04
5	17	3.24	1.61
6	18	2.11	1.99
7	27	3.15	2.21
8	34	4.00	1.53
9	46	1.78	1.67
10	48	2.60	2.18
11	92	3.17	2.15
12	105	1.72	1.65
13	106	2.71	2.12
14	116	4.12	1.59
15	234	2.44	2.17
16	263	1.89	1.82
17	354	4.28	1.60
18	422	3.20	2.26
19	1298	3.78	1.68
20	1828	4.49	1.39
21	2316	3.35	2.18
22	2654	3.74	2.08

We consider many metrics here. The first time we introduce a metric, the name will be in **bold face**.

To address the first requirement, we turn to a metric commonly used to evaluate ACF algorithms, **Mean Absolute Error**, or MAE. This is defined as the average absolute difference between the predicted rating and actual rating over all users and items in a test set. Algorithm performance is evaluated before and after each attempted attack to determine each attack's effect on the system as a whole.

Despite MAE's popularity in ACF algorithm evaluation, there is a mismatch between what it measures and what often matters to users in a recommender system [69]. Recommender systems are a decision support tool, and in general, the accuracy of the exact predicted value is of less importance than whether the specific set of recommendations that are shown to a user are good.

In practice, many recommender systems are used in e-commerce sites to help customers find products to purchase. Many of these systems offer interfaces that produce lists of recommendations for their users, rather than require the user to explicitly state which products she is considering. Schafer et al. found that by far the most common output of e-commerce recommender engines was suggestions of items for the customers to purchase [135]. Therefore, the quality of the *recommendations* should be emphasized rather than the quality of the predictions.

Furthermore, since users do not tend browse very "deeply" when shown a list of things to examine, the items near the top of a recommendation list should be emphasized when evaluating algorithm quality. We analyzed *MovieLens* user behavior and discovered that the median recommendation search session ends within the first 40 items displayed. Figures 7.1 and 7.2 show the browse depth of 137,991 *MovieLens* recommendation (top-N) searches. Figure 7.2 is truncated at 400 items so that the low median is easier to visualize. Figure 7.1 shows the entire set of data on a log-log scale so that the long tail is visible.

The phenomenon that users do not navigate deeply through a list of results may be even more acute in other domains. For instance, one study found that 54 percent of Internet search engine users view only a single page of search results in each session [79]. Developing a top-N recommendation accuracy metric for entire recommender systems is beyond the scope of our work; however, we will develop an attack effectiveness metric that is centered around recommendations rather than predictions.

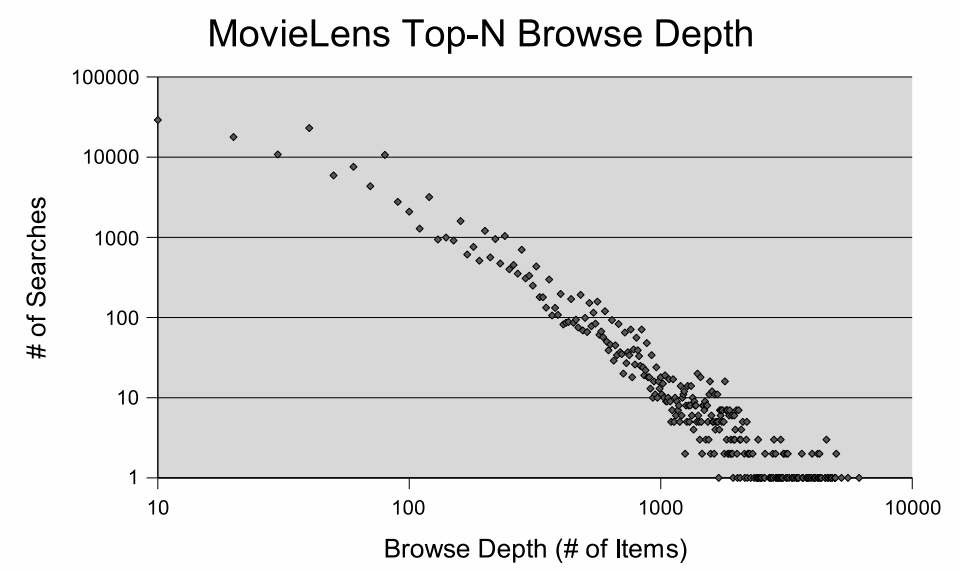

Figure 7.1: Log-log graph of top-*N* search browse depths in *MovieLens*. The median search depth is 40, and 79% of recommendation searches end at or before 80 items.

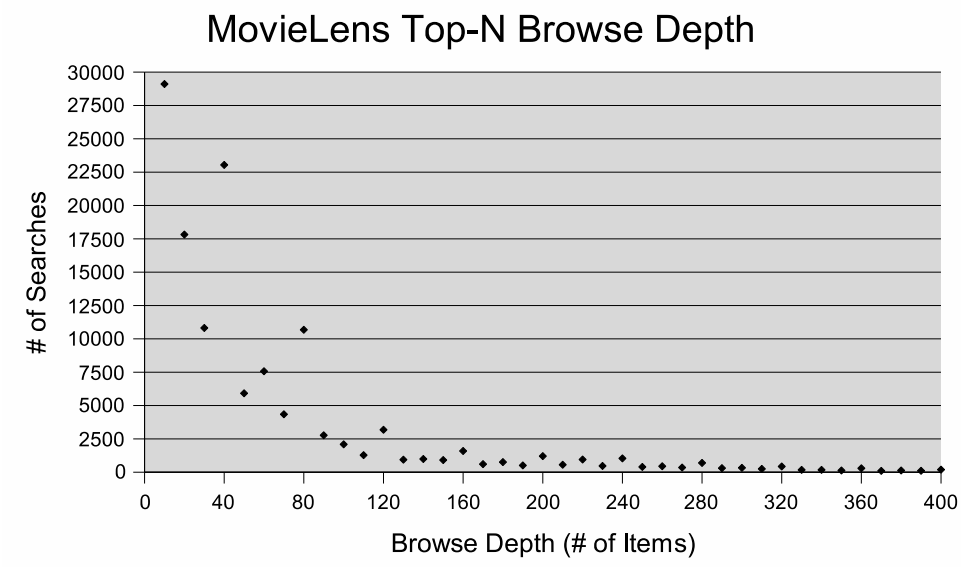

Figure 7.2: Partial graph of top-*N* search browse depths in *MovieLens*. The long tail to the right of 400 items is not shown.

Before we do so, we first examine the attack effectiveness metrics used in previous work. One metric that is introduced in [112] is the **Stability of Prediction** metric, which measures the relative number of predictions for target items that are *not* manipulated beyond some given threshold. This metric does not directly address the notion of manipulating top-N recommendation lists. It treats all prediction manipulations of equal amount as being of equal importance. However, this is intuitively not the case. For instance, on a 5-point rating scale, causing a prediction to change from 2 to 3 is far less meaningful than moving it from 4 to 5. An item with a 3-point prediction is usually far less likely to appear on a recommendation list than an item with a 5-point prediction.

Another metric also introduced in [112] is the **Power of Attack** metric, which is defined as the average change in prediction toward some target value (usually the minimum allowed rating r_{min} or the maximum allowed rating r_{max}) over all target users and items. We believe that this has the same drawbacks as the Stability of Prediction metric. Since Power of Attack is defined *differently* in the authors' earlier work ([113]), we will refer to this metric as **Prediction Shift** when presenting our results.

We will use the original version of the metric defined in [113] when referring to **Power of Attack**. This is defined as the percentage of predictions for target items *not* manipulated to some target value (again, r_{min} or r_{max}). For attacks with a push intent, we find that this metric is somewhat useful for gauging impact on recommendation lists, even though it is a prediction-based metric. An item that has its predicted rating manipulated to r_{max} is indeed likely to appear in a top-N list.

However, Power of Attack falls short in a number of ways. First, the metric is less suitable for measuring the effectiveness of attacks with a nuke intent. In the vast majority of cases, it is unnecessary to manipulate a prediction to r_{min} to cause it to not appear in a top-N. Secondly, some algorithms are quite conservative in making extreme predictions of r_{min} or r_{max}. This metric is less able to accurately gauge attack performance for such cases, as the values the metric takes will be universally high (remember: high metric values mean most items *not* successfully manipulated to the target prediction value).

Conversely, some algorithms may be more liberal in producing extreme predictions. This metric does not take into account the possibility that more than N items can have a predicted rating of r_{max} – in such cases, the top-N recommendation list is not well-defined, and the method used to break such "ties" to determine which N items to display is implementation-dependent.

So, we are left with a need to define a metric that directly measures the effect of an attack on top-N recommendation lists. Furthermore, the metric should take into account the possibility of items tied for inclusion in these lists. Based on these requirements, we propose a metric called **Expected Top-N Occupancy** (ExpTop$N$). The metric is defined as the expected number of occurrences of *target items* in a top-N recommendation list, measured over all users, assuming that the displayed ordering of items tied at any particular rank is random. In the absence of any information about the algorithm implementation, this metric treats all possible tiebreaker functions equally, which has the beneficial side effect of leading to a metric that gives a lower value for attacks that result in many ties involving target items.

For example, let the target items be items E and F, and suppose that the candidates for inclusion in a top-5 recommendation list for some user are as shown below.

Rank	Item	Pred
1	B	5.0
2	E	4.9
3	D	4.7
4	A	4.5
4	C	4.5
4	F	4.5

In this case, there are three items tied for the fourth and fifth entries in the top-5 list. One of these three items is in the target set. Assuming all orderings of these three items are equally likely, the Expected Top-5 Occupancy for this user is $1.\overline{666}$. The target item E in the top-5 contributes 1, and item F being in two out of the three permutations for the fourth and fifth items displayed on the list contributes $0.\overline{666}$.

By examining the change in this metric caused by an attack, one can determine the attack's effect as perceived by users of the recommender system. To reflect actual *MovieLens* usage, we will use $N = 40$ in accordance with our system usage analysis. To compare this metric with the prediction-centric ones, we will report MAE as well as the value of both Power of Attack metrics. The earlier definition from [113] will be referred to as Power of Attack (POA) and the later definition from [112] will be referred to as Prediction Shift (PredShift).

Table 7.2: **Changes in Predictions.** Effect of attacks as measured by the Prediction Shift metric (PredShift) and the change in MAE (ΔMAE). An increase in MAE indicates lower overall predictive accuracy.

Algorithm	Intent	Attack	Bots	PredShift	ΔMAE
User-user	Push	Random	25	0.499	0.002
			50	0.671	0.004
			100	0.830	0.009
		Average	25	1.032	0.006
			50	1.189	0.011
			100	1.300	0.019
	Nuke	Random	25	0.422	0.002
			50	0.589	0.004
			100	0.759	0.010
		Average	25	0.656	0.007
			50	0.815	0.014
			100	0.956	0.023
Item-item	Push	Random	25	0.030	0.002
			50	0.053	0.002
			100	0.069	0.004
		Average	25	0.363	0.002
			50	0.426	0.004
			100	0.471	0.010
	Nuke	Random	25	-0.046	0.002
			50	-0.069	0.002
			100	-0.092	0.004
		Average	25	0.332	0.003
			50	0.354	0.006
			100	0.361	0.014

7.4 Results and Analysis

We now step through each of our hypotheses, stating our results and whether they support or refute the hypothesis.

Hypothesis 1 *Different ACF algorithms respond differently to shilling attacks.*

Predictions. Table 7.2 shows how each of the attacks affected the predictions made by the ACF algorithms. This table reports the two prediction-centric metrics, PredShift and change

Table 7.3: **Changes in Recommendations.** Effect of attacks as measured by the Power of Attack metric (POA) and percent change in Expected Top-40 Occupancy (ExpTop40). For POA, a lower value means the attack was more effective. The value of Expected Top-40 Occupancy pre-attack is 0.57 for the user-user algorithm and 0.24 for item-item.

Algorithm	Intent	Attack	Bots	POA	ExpTop40
User-user	Push	Random	25	0.900	711%
			50	0.865	1190%
			100	0.816	1649%
		Average	25	0.715	1286%
			50	0.609	1674%
			100	0.519	1918%
	Nuke	Random	25	0.943	-39%
			50	0.928	-33%
			100	0.908	-32%
		Average	25	0.963	-67%
			50	0.952	-70%
			100	0.943	-75%
Item-item	Push	Random	25	1.000	150%
			50	1.000	171%
			100	1.000	229%
		Average	25	0.999	158%
			50	0.999	154%
			100	0.999	117%
	Nuke	Random	25	0.954	146%
			50	0.954	204%
			100	0.954	333%
		Average	25	0.955	-33%
			50	0.955	-54%
			100	0.954	-71%

in MAE. These metrics are most appropriate for applications that form predictions for items selected by the user. First, we examine the attack effect as measured by the prediction shift metric. We delay discussion of MAE until the presentation of the hypothesis about detecting shilling, because we found MAE not very useful for the other hypotheses. In general the MAE changes were very small, and seem unlikely to represent noticeable change to users.

The user-user algorithm responds very strongly to all attacks; that is, the attacks are generally successful in manipulating the predictions for items in the target set. For the push attacks,

the PredShift metric indicates that AverageBot-based attacks are able to raise the predictions by over one point, and RandomBot-based attacks are somewhat less effective. A similar pattern is seen for nuke attacks.

On the other hand, the item-item algorithm responds far less strongly. According to PredShift, the *most* effective push attack on item-item (100 AverageBots) has an effect on predictions that is comparable to the *least* effective push attack on user-user (25 RandomBots). RandomBots have an even weaker effect on item-item predictions – no average prediction shift is greater than one-tenth of a point.[5]

Recommendations. Next, we turn to table 7.3, which shows the results of the attacks in terms of metrics that measure their ability to affect recommendations. These metrics are more appropriate for applications that generate lists of suggested items for their users.

According to the POA metric, push attacks on the user-user algorithm are more successful than nuke attacks, with AverageBot being superior to RandomBot. In the MovieLens domain push attacks may be easier because the average rating is higher than the midpoint of the scale. With the item-item algorithm, push attacks seem ineffective – the values of 1 and 0.999 indicate that essentially no predictions for items in the target set were successfully manipulated to the target predicted value of r_{max}, or 5. Nuke attacks on item-item are roughly as effective as they are on user-user, according to POA.

The ExpTop40 metric should be sensitive to prediction changes that influence the top 40 items, whether or not those changes move the predictions to the top of the scale. With this metric, we see a striking difference in response between the two algorithms. The user-user algorithm is profoundly affected by push attacks, with AverageBot again being more successful than RandomBot. Note that a 100-AverageBot attack causes a nineteen-fold increase in how often the target items are recommended in the top 40 items!

Push attacks have a far more subdued effect on the item-item algorithm according to ExpTop40. In the best case, ExpTop40 increases by 229% with a 100-RandomBot push attack. Unexpectedly, we see that a 100-RandomBot *nuke* attack has an even greater effect, causing target items to be recommended over three times as often! While there is a large relative effect on item-item recommendations, the absolute effect is still modest. Even after the 100-RandomBot

[5]Note that using RandomBots in a nuke attack actually *increases* predictions for the item slightly. We are unsure why this happens, though we have been able to show in small-scale examples that attacks on one user in item-item sometimes result in a reverse effect for other users. That is, if an item is pushed for some users, that item is nuked for other users. Perhaps the increase is due to a reversal on a subset of users.

push attack, less than one occurrence of a target item appears in each user's top-40 on average.

Looking at the nuke attacks, we see that the AverageBot-based attacks are successful on both ACF algorithms in manipulating the top-40. With 100 AverageBots, the number of items from the target set appearing in a user's top-40 is reduced by about 75% on the user-user algorithm and 71% on the item-item algorithm.

So, we find that the user-user and item-item algorithms do respond differently, particularly in how their recommendation outputs are affected under push attacks. We judge that the evidence supports the hypothesis that different algorithms respond differently to shilling attacks. ACF algorithms that are not based on the nearest-neighbor approach may exhibit even more variability.

Hypothesis 2 *Shilling attacks affect recommender algorithms differently from prediction algorithms.*

To address this hypothesis, we compare the values of the PredShift, POA, and ExpTop40 metrics. PredShift and POA are prediction metrics, while ExpTop40 is a recommendation metric. For many of the tests both classes of metrics move in the same direction. For instance, under an AverageBot-based push attack on user-user, PredShift is at least a half point, POA shows that at least thirty percent of predictions were successfully manipulated to r_{max}, and the change in ExpTop40 shows that many target items are pushed into the top 40.

However, there are notable differences, too. For instance, The PredShift metric indicates that RandomBot-based push attacks have a nearly negligible effect on item-item's predictions for target items. The ExpTop40 metric, on the other hand, says that up to twice as many target items are in the top 40 after this attack.

Conversely, with Average-bot based push attacks on item-item, PredShift shows that there is an measurable effect on the predictions, but the effect on recommendations as measured by ExpTop40 are minimal. Thus, this attack does appear to have different effects on the prediction and recommendation modes of item-item.

Note that in a few cases, the POA metric shows diametrically opposed results in comparison with the other metrics. In the RandomBot nuke attacks on item-item, PredShift and ExpTop40 show that the predictions and recommendation frequencies for target items *increased*, while POA indicates some success in reducing predicted values for target items to r_{min}, or 1. In fact, POA indicates just as much success for this attack as it does for the AverageBot nuke attack on

item-item, which PredShift and ExpTop40 agree actually does nuke the target items. POA is unable to distinguish between AverageBot and RandomBot attacks on item-item even though each attack has different effects on the predictions and recommendations according to the other metrics.

The high values POA takes for attacks on the item-item algorithm is a result of one of the weaknesses we mentioned earlier. Item-item tends to be more conservative than user-user in yielding very high or very low valued predictions. Thus, it is difficult to manipulate a prediction to either extreme of the scale, and as a result, the POA metric is generally unable to discern the effectiveness of attacks on item-item. Because of this deficiency, we recommend against using POA as the sole metric for evaluating shilling attack performance.

Overall, the evidence for this hypothesis is mixed. Usually, all of the metrics we studied move in the same direction, and it is hard to directly compare them. However, in a few cases there are clear differences between an attack's impact on predictions and on recommendations. We believe these cases are sufficient evidence to argue that selection of an appropriate metric is important. If recommendation is more important than prediction for the recommender system, then a recommendation-based metric such as Expected Top-*N* Occupancy should be used. On the other hand, if prediction-oriented tasks are more important, a prediction-based metric such as Prediction Shift should be used.

Hypothesis 3 *Shilling attacks are not detectable using traditional measures of algorithm performance.*

We return to the last column of table 7.2, which shows the change in MAE caused by each attack. The MAEs are obtained by performing five-fold cross-validation on an 80%/20% test/train split of the data set before and after each attack. Note that no attack increases the MAE by more than 0.023, and that RandomBot attacks on item-item induce extremely small reductions in accuracy. A change in MAE of 0.023 seems small, but is comparable to the improvements in MAE that are considered significant differences among ACF algorithms [67].

As an aside, it is unclear whether *people* are able to perceive such differences in system accuracy. Cosley et al. [30] show that user satisfaction decreases significantly with a recommender that has many of its predictions intentionally shifted by one point on a five-point scale. However, the average effect on accuracy from our shilling attacks is well over an order of magnitude smaller, and is distributed across a large set of items. Furthermore, the system's user

interface may obscure small changes in prediction – for instance, *MovieLens* only displays predictions in half-point increments. Of course, whether users can detect changes in MAE is not directly related to whether MAE can be used by system operators to detect shills.

Overall, we find that insufficient evidence exists to support or reject this hypothesis. Crude attacks may be detectable by watching standard metrics such as MAE, but sophisticated attacks may be effective in achieving their goals, yet remain "under the radar" in terms of their effects on overall recommender system performance. Still, we believe that this ambiguity suggests that other ACF algorithm performance metrics are necessary to reliably detect attacks.

Hypothesis 4 *Ratings distribution of the target item influences attack effectiveness.*

We hypothesized that the properties of an item's ratings distribution has an effect on how much impact an attack has on that item. As mentioned earlier, the properties we are studying are popularity, likability, and entropy. Intuitively, if an item has few ratings (low popularity) and/or has a high spread of ratings (entropy), it should be easier to manipulate the predictions and recommendations for that item because it is more "volatile" in some sense. Likewise, items that are already well-liked should be easy to push, while items that are disliked by many should be easy to nuke.

We examined the effects of these variables on PredShift and ExpTop40. We only considered push attacks for ExpTop40, since so few of our target items were in top 40 lists prior to the attacks, so nuke attacks would have few items to "nuke."

Figures 7.3, 7.4, and 7.5 show several examples of these relationships (or lack thereof). The likability of an item correlates with PredShift on both the user-user and item-item algorithms, but not with the change in ExpTop40 in either algorithm. The popularity of an item correlates highly with the change in ExpTop40 on user-user, but not on item-item. Furthermore, popularity does not correlate well with PredShift. Finally, the entropy of an item's ratings distribution correlates with neither metric and neither algorithm.

The results only indicate partial support for this hypothesis. However, the supported part does carry an important implication for systems that use the user-user algorithm for *recommendations*: items that have low popularity can be an easy push target for attackers. This result is significant because newly added items in a repository have low popularity (i.e., no ratings), and in an e-commerce environment, new items can be natural targets for push attacks since they are

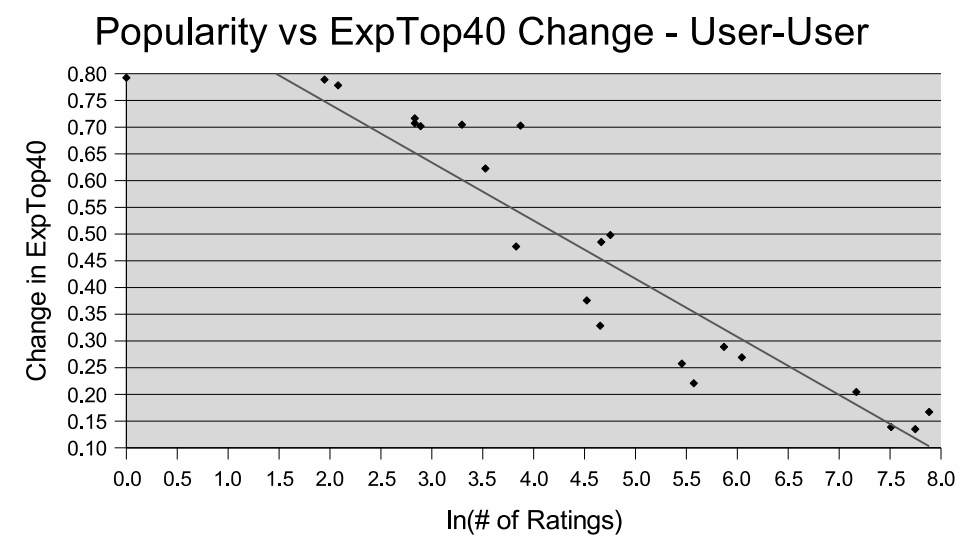

Figure 7.3: Relationship between popularity of an item and the effect of a 100-AverageBot push attack on user-user recommendations ($r^2 = 0.874$)

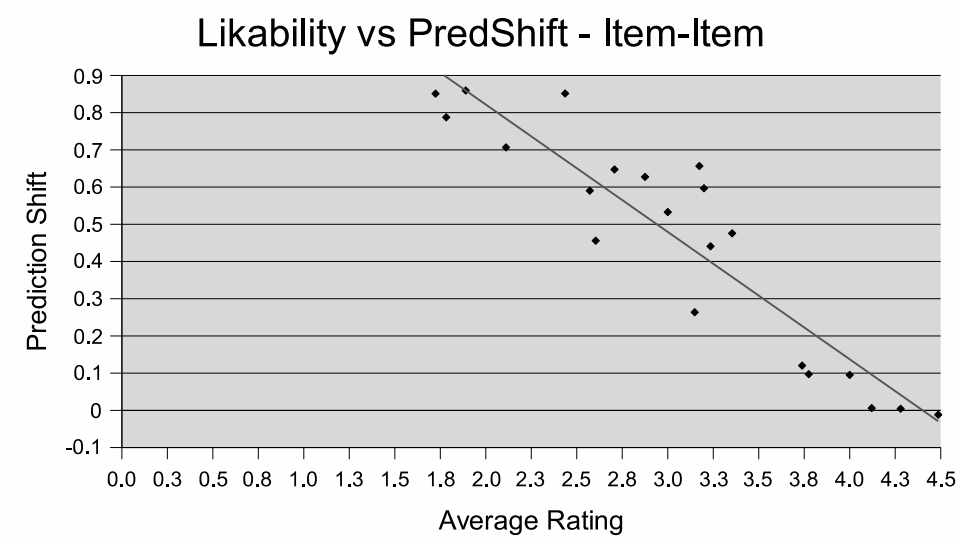

Figure 7.4: Relationship between likability of an item and the effect of a 100-AverageBot push attack on item-item predictions ($r^2 = 0.853$)

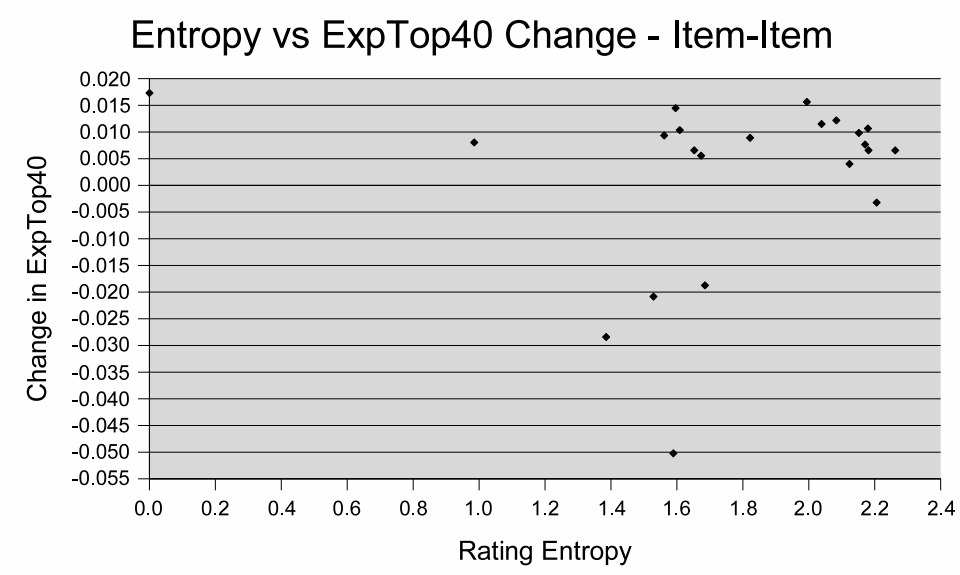

Figure 7.5: (Lack of) Relationship between entropy of an item and the effect of a 100-AverageBot push attack on item-item recommendations ($r^2 = 0.004$)

often newly-released products.[6]

7.5 Discussion

There are some conclusions from this research that can be drawn for operators of recommender systems who would like to reduce the threat of shilling to their systems. We identify those lessons, then close by surveying more recent work in this research area.

Prefer Item-Item. The item-item algorithm was much less affected by the attacks in our study than the user-user algorithm. In many domains, item-item provides recommendations that are the same or better quality as user-user, typically with better runtime performance characteristics. Our study suggests one more reason for operators to prefer item-item: it appears to be more resistant to shilling.

Use Recommendation Metrics. Most recommender systems in practice are used as a source of recommendations, rather than predictions [135]. Our results show that metrics that

[6]In fact, part of the motivation for this study is a long-term MovieLens user who is convinced that new movies are frequently shilled.

are sensitive to changes in prediction accuracy may be less sensitive to changes in recommendation accuracy. We recommend that operators who are producing mostly recommendations use a recommendations-centric metric, such as Expected Top-*N* Occupancy, to evaluate the effectiveness of shilling attacks on their systems. We also recommend that researchers consider focusing on recommendation accuracy, rather than prediction accuracy, in future shilling studies.

Watch Metrics, but Worry Anyway. Operators wish to know whether their recommendations systems are under attack. Watching for sharp changes in the value of traditional algorithm performance metrics such as MAE may be useful for detecting some attacks. However, our results suggest that effective attacks may not be visible through simple aggregate metrics like MAE, so work needs to be done to develop more reliable tools to detect attackers. Note that the attacks we used could be easily detected by watching for individual users with unreasonably large numbers of ratings. Real-world attacks are likely to be more subtle.

Protect New Items. Our experiments show that new or obscure items, particularly in the user-user algorithm, are especially susceptible to attack. Recommender system operators should consider obtaining ratings for such items from a trusted source in order to make them less vulnerable. For instance, in MovieLens, the ratings for a new item could be seeded with ratings from professional critics, with ratings from trusted volunteers, or with filterbots [54].

Related Work. In the years that have passed since we originally conducted this research, the research community has followed up with a rich literature on recommender system and ACF algorithm security.

Some research has focused on developing more sophisticated attack strategies that are effective against more types of ACF algorithms, that require fewer resources or less knowledge to execute, or that are more subtle. For instance, in [109], Mobasher et al. described the *segment attack*, which takes advantage of natural interest segments that occur in many user and item populations (e.g., fans of horror movies, or avid Tom Clancy spy novel readers) in order to produce small attacks that have limited scope, but that are effective even on the item-item algorithm. [126] presents subsequent work that looks at variants of segment attacks that allow for different attack strategies depending on properties of the target item and the ACF algorithm.

Other research has taken the defensive, exploring ways to counteract the effect of shilling attacks. Some techniques approach this as a classification problem, proposing algorithms to distinguish between honest users and shill users by applying heuristics based on empirically observed properties of honest users and known attacks [23, 114, 109, 169]. These techniques

are independent of the ACF algorithm, and are, in some sense, similar to email spam detection.

An alternate defense against shilling attacks is to modify the recommendation process itself to be more robust against attacks. Mobasher et al. developed a hybrid algorithm by introducing content features to the item-item ACF algorithm, and found that the hybrid algorithm was more resistant to attacks [109]. Attacks were unable to modify content-based similarity metrics, and thus had a subdued effect on the hybrid algorithm's recommendations. [130] and [176] propose variants of ACF algorithms that offer mathematically guaranteed levels of resistance against shilling attacks.

In theoretical work, Resnick and Sami prove an information-theoretic bound that establishes a fundamental limitation on efforts to improve ACF algorithm attack resistance [131]. They show that due to the inability to distinguish attackers from legitimate users when little data is available, an inherent tradeoff exists between attack resistance and the amount of genuine ratings information that can be used. That is, in order to achieve a given level of attack resistance, *any* algorithm must also discard a commensurate amount of legitimate information.

The interest driving this body of research forward is indicative of the importance of the challenges that deviant users pose to social production communities. Effective checks and balances are essential when providing users with substantial control over the curation of an information repository.

Chapter 8

Conclusion

In this thesis, we have presented an exploration of two broad areas of collaborative curation in social production communities.

8.1 Theme 1: Collaborative Curation Practices and Mechanisms

Our first research theme focused on how the design of a collaborative curation mechanism can affect the growth and evolution of an SPC information repository. This issue is of substantial interest for SPC designers and operators, since understanding the implications and consequences of different designs can provide them with better control over shaping the growth of the repository in ways that are more closely aligned with the SPC's goals.

Our analysis of Wikipedia's collaborative curation practices showed that in its freeform wiki environment, there appear to be curation dynamics that led to a "top-down" growth pattern favoring development of articles about highly popular topics early in Wikipedia's life. As the encyclopedia has aged, the community's focus has turned toward producing articles about increasingly obscure "long tail" topics. However, we saw evidence of resistance to the arbitrary growth of the tail: there has been an increase in selectivity regarding which topics are deemed acceptable for the encyclopedia. Wikipedia's editors appear quite vigilant in scrutinizing newly-created articles and determining whether each topic meets the established norms regarding inclusion in the encyclopedia. While Wikipedia is a unique and distinctive instance of an SPC whose scale and success may be a once-in-a-generation phenomenon, our findings suggest that interesting self-organized curation practices can emerge in SPCs even when the

system operators impose little structure and establish few rules.

When we turned to studying new-item curation mechanisms in MovieLens, we found interesting tradeoffs between wiki processes and social voting processes. Compared to a wiki process, using a Reddit-like voting process led to the addition of items that carried higher interest from the overall community, as well as a temporary quality advantage. In contrast, the wiki process yielded higher user productivity and faster repository growth. We found no significant differences in user satisfaction between these two mechanisms, so it appears that SPC designers have some latitude in building policies and workflows that meet their goals without risking adverse effects.

In this work, we have studied the curation practices of one large SPC, and have compared two common and simple types of mechanisms, but of course, the design space is vast and many other systems and mechanisms exist. There are numerous design elements and dimensions in SPC curation mechanism design that yield interesting effects on user behavior, user satisfaction, and curation outcomes. Several of these include:

- Explicit **peer review** can be as effective as expert review in some situations [27] as a means of maintaining information quality, but with lower costs.
- Though expensive, **expert review** can perhaps be applied selectively to spot-check contribution validity and quality where it is needed most.
- **Reputation systems** that score users' trustworthiness, contribution value, and/or expertise could be used to help determine which actions to take for a particular submission, who to confer curation privileges to, and so forth.
- **Gamification or competition** (e.g., badges, achievements, rankings, leaderboards, contests) can be a motivational device as well as a tool to guide users toward areas of the SPC repository where there is need for additional work.
- Establishing **roles or hierarchies** within a community may be useful for SPCs that desire more structured processes or explicit delineation of curation responsibilities.

Besides these possibilities, there may even exist as yet unknown mechanisms and design elements that can help mitigate or even eliminate the need for the tradeoffs that we identified.

There are also many other open questions in this space that are ripe for study. Below, we discuss two such questions.

What mechanisms are appropriate for SPCs where the operators require more substantial levels of quality control? Thus far, we have focused on SPCs like Wikipedia that accept the possibility of incorrect or invalid information appearing in the repository and perhaps lingering unnoticed for lengthy periods of time. Not all SPCs can be so tolerant. For example, in citizen science systems, the presence of flawed data submissions can pose a real risk to the success of a project, so data quality is a pressing concern [167, 140]. A cautionary tale to consider here is Larry Sanger's pursuit of quality in Nupedia. Though intricate and exhaustive collaborative curation workflows that require the involvement of domain experts are perhaps effective in theory, they may not be the right answer in practice. Nupedia's article review process apparently imposed too much formality to gather a critical mass. However, perhaps there is a middle ground that strikes a better balance between quality control and user acceptance.

How should curation mechanisms evolve along with the SPC through its lifecycle? The needs and goals of an SPC may change substantially as it matures, and therefore, the "ideal" curation mechanism may be a moving target. In [52], Goldman makes a provocative argument that Wikipedia may need to move away from its wiki model of "free editability" in order to address and overcome the challenges it is facing due to a declining editor population. He notes that there is already evidence that Wikipedia has been moving in that direction with the various editing limitations that have been placed on anonymous and new editors. Goldman speculates that without further changes, the proliferation of non-constructive users will eventually exhaust and overwhelm Wikipedia's community of editors (we note that more recent explorations of Wikipedia suggest that some of these pressing issues are being addressed in part through the use of *bots* and *cyborgs*, automated and semi-automated editing tools that allow for more efficient handling of mundane labor-intensive tasks like cleaning up vandalism [46, 59]).

8.2 Theme 2: Collaborative Curation Challenges

In our second research theme, we looked at three challenges that real-world SPCs face. We focused specifically on studying various properties of SPC users including experience level, demographics, and behaviors, and the effects that abnormal or unexpected skews in these properties have on collaborative curation and the evolution of the SPC information repository. Though

there are myriad challenges that SPCs can face including technical, legal, and governance issues, we chose to focus on users and user behaviors because challenges involving users tend to be omnipresent throughout the life of most SPCs.

We first looked at challenges around group composition, both at a small-scale working group level, and at a large-scale community-wide level. We found that on Wikipedia, the quality of curation decisions made by small working groups is significantly associated with several group composition factors. For instance, groups that are very small or that have a high concentration of newcomers tend to make worse decisions, while groups with high tenure diversity make better decisions. Some of our results are in line with well-known theories and outcomes from social psychology, while others are not. It appears that the asynchronous text-based communications used by many SPCs lead to group interaction dynamics that cannot always be predicted and explained by classic theories developed from observing face-to-face interactions.

Next, we explored the substantial community-wide gender disparity amongst Wikipedia's editors, and the effects that it has had on the encyclopedia's curation. Despite comprising about half of Wikipedia's readership, females account for just one of every six editors, and one of every eleven edits. Furthermore, Wikipedia's male-skewed curators have apparently focused their efforts substantially more on male-interest topics than female-topics: we found a significant gender-oriented deficiency in the encyclopedia's coverage.

We note that such phenomena are not necessarily harmful. Numerous demographic skews exist in other activities. For example, the online game World of Warcraft is dominated by males [175], and golf enthusiasts tend to skew older. Such skews are, however, generally accepted as benign.

On the other hand, the gender-related disparities we observed in Wikipedia, and the other topical coverage skews that have been reported in the literature [65, 84, 57], are problematic for Wikipedia and its ability to achieve its purpose. The Wikimedia Foundation's vision for its endeavors is "a world in which every single human being can freely share in the sum of all knowledge." Our findings suggest that at least in the English-language Wikipedia, knowledge from half of the population is not represented commensurately.

There are other potentially adverse effects as well. Hecht and Gergle note that algorithms and other technologies are increasingly turning to SPCs as a source of information, and posit that the segregation of populations into different SPCs can lead to those technolgoies behaving very differently depending on which information resource they use [65]. They demonstrate

this phenomenon with Explicit Semantic Analysis (ESA), an NLP technique for measuring the semantic relatedness between arbitrary concepts. Their experimentation shows that ESA yields very different results depending on which language Wikipedia is used as the input source of world knowledge. Thus, any applications that utilize ESA may be affected substantially depending on which data source is used ([65] provides one example of such an application: a conversation topic clustering visualization that uses ESA to measure topic similarity).

The issue of how best to address group composition skews in SPCs remains as an open question, especially for large-scale imbalances that are unlikely to be alleviated through known technical approaches like intelligent task routing [29]. An added challenge here is that the root issues may not be solvable through technological innovation. For instance, both our work and other emerging research [26] indicate that an important factor in Wikipedia's gender imbalance is its culture, which is sometimes characterized as contentious, critical, and adversarial. This suggests that an effective path to address the imbalance will likely require social changes in addition to technical modifications.

If nothing is done to control a population-level disparity, it seems plausible that the skew will perpetuate itself, as has been the case with Wikipedia's gender imbalance. To make matters worse, it is possible that balkanization, homophily, and social categorization effects will favor overrepresented subgroups and allow them to control community dynamics, which can serve to further *widen* the skew and excaberate its effects in a vicious cycle.

Finally, we looked at the effects of deviant users who try to manipulate SPCs for their own individual gain. We found that recommender systems, which are commonly found in SPCs used to help users navigate large information repositories, are indeed vulnerable to relatively simple attacks designed to cause specific items to appear more (or less) often in lists of recommended items. Our research in this area helped pioneer a rich field of study in recommender system security, which has led to the development of many new theories and techniques for detecting, mitigating, and limiting the effects of deviance in recommender systems.

Looking forward, we believe that SPCs will continue to be an effective and common approach for creating information repositories. Furthermore, we believe SPCs will offer unprecedented opportunity for groundbreaking interdisciplinary research in computer science, social psychology, economics, organizational studies, and many other fields. The large-scale collaborations that occur in SPCs offer unique settings for studies that might otherwise be infeasible in traditional real-world laboratory or field studies. Our contributions in this thesis have shed light

on several interesting aspects of collaborative curation mechanism design in SPCs, and the challenges that SPCs face as their users work together to construct useful information repositories. We hope that fellow researchers will be able to build upon our work, and that practitioners can benefit from our findings and be able to build and operate more successful SPCs in the future.

References

[1] Nupedia.com Editorial Policy Guidelines. http://web.archive.org/web/20030810192540/http://www.nupedia.com/policy.shtml, July 2001.

[2] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *Proc. WSDM 2008*, pages 207–218, Palo Alto, CA, USA, 2008. ACM.

[3] C. Anderson. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion, July 2006.

[4] O. Arazy and A. Croitoru. The sustainability of corporate wikis: A time-series analysis of activity patterns. *ACM Trans. Manage. Inf. Syst.*, 1(1):6:1–6:24, Dec. 2010.

[5] E. Aronson and J. Mills. The effect of severity of initiation on liking for a group. *The Journal of Abnormal and Social Psychology*, 59(2):177–181, 1959.

[6] S. E. Asch. Effects of group pressure upon the modification and distortion of judgements. *Groups, Leadership, and Men*, pages 177–190, 1951.

[7] A. L. Association. 2011 State of America's Libraries Report. http://www.ala.org/news/mediapresscenter/americaslibraries, Apr. 2011.

[8] B. B. Baltes, M. W. Dickson, M. P. Sherman, C. C. Bauer, and J. S. LaGanke. Computer-Mediated communication and group decision making: A Meta-Analysis. *Organ Behav Hum Dec*, 87(1):156–179, 2002.

[9] R. Bartle. Hearts, clubs, diamonds, spades: Players who suit MUDs. *Journal of MUD research*, 1(1):19, 1996.

[10] A. G. Bedeian and K. W. Mossholder. On the use of the coefficient of variation as a measure of diversity. *Organizational Research Methods*, 3(3):285–297, July 2000.

[11] G. Beenen, K. Ling, X. Wang, K. Chang, D. Frankowski, P. Resnick, and R. E. Kraut. Using social psychology to motivate contributions to online communities. In *Proc. CSCW2004*, Chicago, IL, 2004.

[12] I. Beschastnikh, T. Kriplean, and D. W. McDonald. Wikipedian self-governance in action: Motivating the policy lens. In *Proc. ICWSM 2008*, Seattle, WA, USA, 2008. AAAI.

[13] B. A. Bettencourt and N. Miller. Gender differences in aggression as a function of provocation: A meta-analysis. *Psychol Bull*, 119(3):422–447, May 1996.

[14] J. E. Blumenstock. Size matters: Word count as a measure of quality on Wikipedia. In *Proc. WWW 2008*. ACM.

[15] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 43–52, July 1998.

[16] A. Bruckman, P. Curtis, C. Figallo, and B. Laurel. Approaches to managing deviant behavior in virtual communities. In C. Plaisant, editor, *CHI Conference Companion*, pages 183–184. ACM, 1994.

[17] S. L. Bryant, A. Forte, and A. Bruckman. Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Proc. GROUP 2005*, pages 1–10, Sanibel Island, FL, USA, 2005. ACM.

[18] M. Bunz. YouTube faces 4chan porn attack. *The Guardian*, Jan. 2010.

[19] M. Burke and R. Kraut. Mopping up: Modeling Wikipedia promotion decisions. In *Proc. CSCW 2008*, San Diego, CA. ACM.

[20] E. S. Callahan and S. C. Herring. Cultural bias in Wikipedia content on famous persons. *J. Am. Soc. Inf. Sci. Technol.*, 62(10):1899–1915, Oct. 2011.

[21] J. Canny. Collaborative filtering with privacy via factor analysis. In *SIGIR '02: Proceedings of the 25th International ACM Conference on Research and Development in Information Retrieval*, pages 238–245, Tampere, Finland, 2002. ACM Press.

[22] J. Chen, Y. Ren, and J. Riedl. The effects of diversity on group productivity and member withdrawal in online volunteer groups. In *Proc. CHI 2010*, Atlanta, GA. ACM.

[23] P. Chirita, W. Nejdl, and C. Zamfir. Preventing shilling attacks in online recommender systems. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, WIDM '05, pages 67–74, New York, NY, USA, 2005. ACM.

[24] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *0706.1062*, June 2007.

[25] N. Cohen. Define gender gap? Look up Wikipedia's contributor list. *The New York Times*, Jan. 2011.

[26] B. Collier and J. Bear. Conflict, criticism, or confidence: An empirical examination of the gender gap in Wikipedia contributions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 383–392, New York, NY, USA, 2012. ACM.

[27] D. Cosley, D. Frankowski, S. Kiesler, L. Terveen, and J. Riedl. How oversight improves member-maintained communities. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 11–20, Portland, Oregon, USA, 2005. ACM.

[28] D. Cosley, D. Frankowski, L. Terveen, and J. Riedl. Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In *Proc. CHI 2006*, pages 1037–1046, Montréal, Québec, Canada, 2006. ACM.

[29] D. Cosley, D. Frankowski, L. Terveen, and J. Riedl. SuggestBot: Using intelligent task routing to help people find work in Wikipedia. In *Proc. IUI 2007*, pages 32–41, Honolulu, HI, USA, 2007. ACM.

[30] D. Cosley, S. K. Lam, I. Albert, J. Konstan, and J. Riedl. Is seeing believing? How recommender system interfaces affect users' opinions. In *CHI*, 2003.

[31] D. Cox and D. Oakes. *Analysis of Survival Data*. Chapman and Hall/CRC, June 1984.

[32] C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *ACM Conference on Electronic Commerce*, pages 150–157, 2000.

[33] J. Dibbell. *My Tiny Life: Crime and Passion in a Virtual World*. Holt Paperbacks, 1st edition, Jan. 1999.

[34] J. Donath, P. Kollock, and M. Smith. Identity and deception in the virtual community. In *Communities in Cyberspace*. Routledge, 1999.

[35] W. Dong and W. Fu. Cultural difference in image tagging. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 981–984, New York, NY, USA, 2010. ACM.

[36] S. Drenner, S. Sen, and L. Terveen. Crafting the initial user experience to achieve community goals. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 187–194, Lausanne, Switzerland, 2008. ACM.

[37] R. English and C. M. Schweik. Identifying success and tragedy of FLOSS commons: A preliminary classification of Sourceforge.net projects. In *First International Workshop on Emerging Trends in FLOSS Research and Development, 2007. FLOSS '07*. IEEE, May 2007.

[38] A. Feingold. Gender differences in personality: A meta-analysis. *Psychol Bull*, 116(3):429–456, Nov. 1994.

[39] R. T. Fielding. Shared leadership in the apache project. *Commun. ACM*, 42(4):42–43, Apr. 1999.

[40] A. Forte and A. Bruckman. Scaling consensus: Increasing decentralization in Wikipedia governance. In *Proc. HICSS 2008*, page 157. IEEE Computer Society, 2008.

[41] W. Foundation. Wikistats: Wikimedia Statistics. `http://stats.wikimedia.org/EN/TablesPageViewsMonthlyCombined.htm`, 2012. Accessed: February 11, 2012.

[42] S. Frederick, G. Loewenstein, and T. Odonoghue. Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40(2):351–401, June 2002.

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

[44] D. Fudenberg and J. Tirole. Bayesian games and mechanism design. In *Game Theory*, pages 243–318. The MIT Press, Aug. 1991.

[45] R. S. Geiger and D. Ribes. The work of sustaining order in Wikipedia: The banning of a vandal. In *Proc. CSCW 2010*, pages 117–126, Savannah, GA, USA, 2010. ACM.

[46] S. Geiger. The lives of bots. In G. Lovink and N. Tkacz, editors, *Critical Point of View: A Wikpedia Reader*, pages 78–93. Institute of Network Cultures, Amsterdam, 2011.

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

[48] R. Glott, P. Schmidt, and R. Ghosh. Wikipedia survey – overview of results. Technical report, United Nations University MERIT, Mar. 2010.

[49] D. Goldberg, D. Nichols, B. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[50] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[51] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208, 2006.

[52] E. Goldman. Wikipedia's Labor Squeeze and its Consequences. *Journal of Telecommunications and High Technology Law, Vol. 8*, 2009.

[53] M. L. Goldstein, S. A. Morris, and G. G. Yen. Problems with fitting to the power-law distribution. *The European Physical Journal B - Condensed Matter and Complex Systems*, 41(2):255–258, 2004.

[54] N. Good, B. Schafer, J. Konstan, A. Borchers, B. Sarwar, J. Herlocker, and J. Riedl. Combining collaborative filtering with personal agents for better recommendations. In *Proceedings of the 1999 Conference of the American Association of Artificial Intelligence (AAAI-99)*, July 1999.

[55] A. G. Greenwald and M. R. Banaji. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4–27, 1995. PMID: 7878162.

[56] J. R. Hackman and N. Katz. Group behavior and performance. In *Handbook of Social Psychology*, volume 2, pages 1208–1251. Wiley, New York, 5th edition, 2010.

[57] A. Halavais and D. Lackaff. An analysis of topical coverage of Wikipedia. *Journal of ComputerMediated Communication*, 13(2):429–440, Jan. 2008.

[58] A. Halfaker, A. Kittur, and J. Riedl. Don't bite the newbies: How reverts affect the quantity and quality of Wikipedia work. In *Proc. WikiSym 2011*, WikiSym '11, pages 163–172, New York, NY, USA, 2011. ACM.

[59] A. Halfaker and J. Riedl. Bots and cyborgs: Wikipedia's immune system. *Computer*, 45(3):79–82, Mar. 2012.

[60] J. Halliday. Digg investigates claims of conservative 'censorship'. *The Guardian*, Aug. 2010.

[61] G. Hardin. The tragedy of the commons. *Science*, 162:1243–1248, 1968.

[62] F. M. Harper, D. Frankowski, S. Drenner, Y. Ren, S. Kiesler, L. Terveen, R. Kraut, and J. Riedl. Talk amongst yourselves: Inviting users to participate in online conversations. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 62–71, Honolulu, Hawaii, USA, 2007. ACM.

[63] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan. Predictors of answer quality in online Q&A sites. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 865–874, Florence, Italy, 2008. ACM.

[64] B. Heater. 4Chan followers hack time's 'Influential' poll. *PCMAG*, Apr. 2009.

[65] B. Hecht and D. Gergle. The tower of Babel meets web 2.0: User-generated content and its applications in a multilingual context. In *Proc. CHI 2010*, CHI '10, pages 291–300, New York, NY, USA, 2010. ACM.

[66] B. J. Hecht and D. Gergle. On the "localness" of user-generated content. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 229–232, New York, NY, USA, 2010. ACM.

[67] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 1999 Conference on Research and Development in Information Retrieval (SIGIR-99)*, Aug. 1999.

[68] J. Herlocker, J. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2000. CHI Letters 5(1).

[69] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, Jan. 2004.

[70] S. Herring. Computer-mediated communication on the internet. *Annual review of information science and technology*, 36(1):109–168, 2002.

[71] S. C. Herring. Gender and power in on-line communication. In J. Holmes and M. Meyerhoff, editors, *The Handbook of Language and Gender*, pages 202–228. Blackwell, 2003.

[72] R. R. Hocking. *Methods and applications of linear models*. John Wiley and Sons, Mar. 2003.

[73] T. Hogg and K. Lerman. Stochastic models of User-Contributory web sites. In *Proc. ICWSM 2009*, Mar. 2009.

[74] L. Hong, G. Convertino, and E. H. Chi. Language matters in Twitter: A large scale study. In *Fifth International AAAI Conference on Weblogs and Social Media*, May 2011.

[75] M. G. Hoy and G. Milne. Gender differences in privacy-related measures for young adult Facebook users. *Journal of Interactive Advertising*, 10(2):28–46, 2010.

[76] F. Hunt and P. Johnson. On the Pareto distribution of SourceForge projects. In *Proceedings of the Open Source Software Development Workshop*, pages 122–129, 2002.

[77] Information Solutions Group. 2010 social gaming research. Technical report, 2010.

[78] I. Janis. *Groupthink*. Houghton Mifflin Boston, 1982.

[79] B. J. Jansen and A. Spink. An analysis of web documents retrieved and viewed. In *Internet Computing Conference*, Las Vegas, 2003.

[80] S. J. Karau and K. D. Williams. Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65(4):681–706, 1993.

[81] A. Keen. *The Cult of the Amateur: How Today's Internet is Killing Our Culture*. Crown Business, June 2007.

[82] S. Kiesler, J. Siegel, and T. W. McGuire. Social psychological aspects of computer-mediated communication. *American Psychologist*, 39(10):1123–1134, 1984.

[83] A. Kittur, E. H. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Proc. CHI 2007*, Montreal, Quebec, Canada, 2007. ACM.

[84] A. Kittur, E. H. Chi, and B. Suh. What's in Wikipedia?: Mapping topics and conflict using socially annotated category structure. In *Proc. CHI 2009*, CHI '09, pages 1509–1512, New York, NY, USA, 2009. ACM.

[85] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi. He says, she says: Conflict and coordination in Wikipedia. In *Proc. CHI 2007*, San Jose, CA. ACM.

[86] T. Kriplean, I. Beschastnikh, D. W. McDonald, and S. A. Golder. Community, consensus, coercion, control: CS*W or how policy mediates mass participation. In *Proc. GROUP 2007*, pages 167–176, Sanibel Island, FL, USA, 2007. ACM.

[87] S. K. Lam, J. Karim, and J. Riedl. The effects of group composition on decision quality in a social production community. In *Proc. GROUP 2010*, Sanibel Is., FL. ACM.

[88] S. K. Lam and J. Riedl. Shilling recommender systems for fun and profit. In *Proceedings of the 13th international conference on World Wide Web*, pages 393–402, New York, NY, USA, 2004. ACM.

[89] S. K. Lam and J. Riedl. Is Wikipedia growing a longer tail? In *Proc. GROUP 2009*, pages 105–114, Sanibel Island, FL, USA, 2009. ACM.

[90] S. K. Lam, A. Uduwage, Z. Dong, S. Sen, D. R. Musicant, L. Terveen, and J. Riedl. WP:Clubhouse?: An exploration of Wikipedia's gender imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, pages 1–10, New York, NY, USA, 2011. ACM.

[91] C. Lampe and P. Resnick. Slash(dot) and burn: Distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 543–550, Vienna, Austria, 2004. ACM.

[92] J. Lanier. *You Are Not a Gadget: A Manifesto*. Knopf, Jan. 2010.

[93] A. Lenhart, K. Purcell, A. Smith, and K. Zickuhr. Social media and young adults. `http://pewinternet.org/Reports/2010/Social-Media-and-Young-Adults.aspx`, February 2010. Accessed March 19, 2011.

[94] J. M. Levine and R. L. Moreland. Progress in small group research. *Annu Rev Psychol*, 41(1):585–634, 1990.

[95] M. A. Lieberman, A. Wizlenberg, M. Golant, and M. Di Minno. The impact of group composition on Internet support groups: Homogeneous versus heterogeneous parkinson's groups. *Group Dynamics: Theory, Research, and Practice*, 9(4):239–250, 2005.

[96] S. Lim and N. Kwon. Gender differences in information behavior concerning Wikipedia, an unorthodox information source? *Library & Information Science Research*, 32(3):212–220, 2010.

[97] Y. Liu and E. Agichtein. You've got answers: Towards personalized models for predicting success in community question answering. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 97–100, Columbus, Ohio, 2008. Association for Computational Linguistics.

[98] T. Locher, P. Moor, S. Schmid, and R. Wattenhofer. Free riding in BitTorrent is cheap. *Proceedings of the 2006 Workshop on Hot Topics in Networks*, 2006.

[99] E. A. Locke, G. P. Latham, K. J. Smith, and R. E. Wood. *A Theory of Goal Setting & Task Performance*. Prentice Hall College Div, Jan. 1990.

[100] P. J. Ludford, D. Cosley, D. Frankowski, and L. Terveen. Think different: Increasing online community participation using uniqueness and group dissimilarity. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 631–638, Vienna, Austria, 2004. ACM.

[101] J. Margolis and A. Fisher. *Unlocking the clubhouse: Women in computing*. MIT Press, 2002.

[102] S. Mcnee, S. Lam, C. Guetzlaff, J. Konstan, J. Riedl, M. Rauterberg, M. Menozzi, J. Wesson, M. Rauterberg, M. Menozzi, and J. Wesson. Confidence displays and training in recommender systems. In *INTERACT*. IOS Press, 2003.

[103] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annu Rev Sociol*, 27(1):415–444, 2001.

[104] B. Miller, S. K. Lam, I. Albert, J. Konstan, and J. Riedl. Movielens unplugged: Experiences with a disconnected recommender system. In *Proceedings of the Intelligent User Interface Conference*, 2003.

[105] B. Miller, J. Riedl, and J. Konstan. GroupLens for Usenet: Experiences in applying collaborative filtering to a social information system. In C. Leug and D. Fisher, editors, *From Usenet to CoWebs: Interacting with Social Information Spaces*. Springer-Verlag, 2002.

[106] B. N. Miller. *Toward a Personal Recommender System*. PhD thesis, University of Minnesota, 2002.

[107] E. Mills. Study: eBay sellers gaming the reputation system? *CNET*, Jan. 2007.

[108] D. Milne, O. Medelyan, and I. H. Witten. Mining domain-specific thesauri from Wikipedia: A case study. In *Proc. WI 2006*, pages 442–448. IEEE CS, 2006.

[109] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Trans. Internet Technol.*, 7(4):23, 2007.

[110] C. Nass and Y. Moon. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, pages 81–103, 2000.

[111] A. A. of Museums Curators Committee. A Code of Ethics for Curators. `http://www.curcom.org/ethics.php`, 2009.

[112] M. P. O'Mahony, N. Hurley, N. Kushmerick, and G. Silvestre. Collaborative recommendation: A robustness analysis. *ACM Transactions on Internet Technology*, 2003. Special Issue on Machine Learning for the Internet.

[113] M. P. O'Mahony, N. J. Hurley, and G. C. Silvestre. Promoting recommendations: An attack on collaborative filtering. In *Proceedings of the 13th International Conference on Database and Expert Systems Applications*, pages 494–503. Springer Verlag, 2002.

[114] M. P. O'Mahony, N. J. Hurley, and G. C. Silvestre. Detecting noise in recommender system databases. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 109–115, Sydney, Australia, 2006. ACM.

[115] K. Panciera, R. Priedhorsky, T. Erickson, and L. Terveen. Lurking? Cyclopaths?: A quantitative lifecycle analysis of user behavior in a geowiki. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1917–1926, New York, NY, USA, 2010. ACM.

[116] N. B. Peddibhotla and M. R. Subramani. Contributing to public document repositories: A critical mass theory perspective. *Organization Studies*, 28(3):327–346, Mar. 2007.

[117] Pew Research Center. Internet Usage Over Time Trends Dataset. `http://www.pewinternet.org/Static-Pages/Trend-Data/Usage-Over-Time.aspx`, November 2010. Accessed March 19, 2011.

[118] M. Piatek, T. Isdal, T. Anderson, A. Krishnamurthy, and A. Venkataramani. Do incentives build robustness in BitTorrent. In *Proceedings of NSDI 2007*, 2007.

[119] D. Pogue. Belkin employee paid users for good reviews. *Pogue's Posts Blog (New York Times)*, Jan. 2009.

[120] E. A. Posner. Does political bias in the judiciary matter?: Implications of judicial bias studies for legal and constitutional reform. *U Chi L Rev*, 75(2):853–883, 2008.

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

[122] R. Priedhorsky, J. Chen, S. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in Wikipedia. In *Proc. GROUP 2007*, Sanibel Is., FL. ACM.

[123] R. Priedhorsky, M. Masli, and L. Terveen. Eliciting and focusing geographic volunteer work. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 61–70, Savannah, Georgia, USA, 2010. ACM.

[124] A. J. Quinn and B. B. Bederson. Human computation: A survey and taxonomy of a growing field. In *Proc. CHI 2011*, CHI '11, pages 1403–1412, New York, NY, USA, 2011. ACM.

[125] A. M. Rashid, K. Ling, R. D. Tassone, P. Resnick, R. Kraut, and J. Riedl. Motivating participation by displaying the value of contribution. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 955–958, Montréal, Québec, Canada, 2006. ACM.

[126] S. Ray and A. Mahanti. Filler item strategies for shilling attacks against recommender systems. In *Hawaii International Conference on System Sciences*, Los Alamitos, CA, USA, 2009. IEEE Computer Society.

[127] J. Reagle and L. Rhue. Gender bias in Wikipedia and Britannica. *Communication Studies Faculty Publications*, Jan. 2011.

[128] Y. Ren, R. Kraut, and S. Kiesler. Applying common identity and bond theory to design of online communities. *Organization Studies*, 28(3):377–408, Mar. 2007.

[129] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In *CSCW '94: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pages 175–186, Chapel Hill, North Carolina, United States, 1994. ACM Press.

[130] P. Resnick and R. Sami. The influence limiter: Provably manipulation-resistant recommender systems. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 25–32, Minneapolis, MN, USA, 2007. ACM.

[131] P. Resnick and R. Sami. The information cost of manipulation-resistance in recommender systems. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 147–154, Lausanne, Switzerland, 2008. ACM.

[132] L. Sanger. The early history of Nupedia and Wikipedia: A memoir. In C. DiBona, M. Stone, and D. Cooper, editors, *Open Sources 2.0: The Continuing Evolution*. O'Reilly Media, 1st edition, Oct. 2005.

[133] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW '01: Proceedings of the 10th International Conference on World Wide Web*, pages 285–295, Hong Kong, 2001. ACM Press.

[134] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. Application of dimensionality reduction in recommender system – a case study. In *ACM WebKDD 2000 Web Mining for E-Commerce Workshop*, Boston, MA, USA, 2000.

[135] J. Schafer, J. Konstan, and J. Riedl. Electronic commerce recommender applications. *Data Mining and Knowledge Discovery*, Jan. 2001.

[136] S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. tagging, communities, vocabulary, evolution. In *Proceedings of the ACM 2006 Conference on CSCW*, Banff, Alberta, Canada, 2006.

[137] S. Sen, J. Vig, and J. Riedl. Learning to recognize valuable tags. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 87–96, Sanibel Island, Florida, USA, 2009. ACM.

[138] C. Shah and J. Pomerantz. Evaluating and predicting answer quality in community QA. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 411–418, Geneva, Switzerland, 2010. ACM.

[139] U. Shardanand and P. Maes. Social information filtering: Algorithms for automating 'word of mouth'. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 210–217, 1995.

[140] S. A. Sheppard and L. Terveen. Quality is a verb: The operationalization of data quality in a citizen science community. In *Proceedings of the 7th International Symposium on*

Wikis and Open Collaboration, WikiSym '11, pages 29–38, New York, NY, USA, 2011. ACM.

[141] C. Shirky. A group is its own worst enemy. In *The Best Software Writing I*, pages 183–209. Apress, 2005.

[142] J. Siegel, V. Dubrovsky, S. Kiesler, and T. W. McGuire. Group processes in computer-mediated communication. *Organ Behav Hum Dec*, 37(2):157–187, 1986.

[143] A. Smith and L. Rainie. 8% of online Americans use Twitter. `http://www.pewinternet.org/Reports/2010/Twitter-Update-2010/Findings.aspx`, December 2010. Accessed March 19, 2011.

[144] I. Steiner. *Group process and productivity*. Academic Press, New York, 1972.

[145] B. Suh, E. H. Chi, B. A. Pendleton, and A. Kittur. Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations. In *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, pages 163–170. IEEE Computer Society, 2007.

[146] B. Suh, G. Convertino, E. H. Chi, and P. Pirolli. The singularity is not near: Slowing growth of Wikipedia. In *Proc. WikiSym 2009*, Orlando, FL. ACM.

[147] J. Suler. The online disinhibition effect. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, 7(3):321–326, June 2004. PMID: 15257832.

[148] N. Summers. Walking the cyberbeat. *Newsweek Magazine*, May 2009.

[149] J. Surowiecki. *The Wisdom of Crowds*. Anchor, Aug. 2005.

[150] K. Sweet. 57M U.S. consumers playing social network games: Survey. *Fox Business*, August 23 2010.

[151] H. Taniguchi. Men's and women's volunteering: Gender differences in the effects of employment and family characteristics. *Nonprofit and Voluntary Sector Quarterly*, 35(1):83–101, 2006.

[152] The Linux Foundation. Linux Adoption Trends 2012: A Survey of Enterprise End Users. Technical report, Jan. 2012.

[153] THEwikiStics. Yearly wikimedia page hits comparison. `http://wikistics.falsikon.de/2008/`, 2008.

[154] B. K. Thorn and T. Connolly. Discretionary data bases. *Communication Research*, 14(5):512–528, Oct. 1987.

[155] Z. Tufekci. Grooming, Gossip, Facebook and MySpace – what can we learn about these sites from those who won't assimilate? *Information, Communication & Society*, 11(4):544–564, 2008.

[156] D. van Knippenberg, C. K. W. D. Dreu, and A. C. Homan. Work group diversity and group performance: An integrative model and research agenda. *The Journal of Applied Psychology*, 89(6):1008–1022, Dec. 2004. PMID: 15584838.

[157] V. Venkatesh, M. Morris, and P. Ackerman. A longitudinal field investigation of gender differences in individual technology adoption decision-making processes. *Organizational Behavior and Human Decision Processes*, 83(1):33–60, 2000.

[158] F. Viegas, M. Wattenberg, and M. McKeon. The hidden order of Wikipedia. In *Proc. OCSC 2007*, pages 445–454. 2007.

[159] F. B. Viegas and M. Smith. Newsgroup crowds and AuthorLines: Visualizing the activity of individuals in conversational cyberspaces. In *Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4 - Volume 4*, HICSS '04, Washington, DC, USA, 2004. IEEE Computer Society.

[160] F. B. Viegas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proc. CHI 2004*, pages 575–582, Vienna, Austria, 2004. ACM.

[161] J. Vig, S. Sen, and J. Riedl. Navigating the tag genome. In *Proceedings of the 16th international conference on Intelligent user interfaces*, IUI '11, pages 93–102, New York, NY, USA, 2011. ACM.

[162] L. Von Ahn. *Human computation*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2005. AAI3205378.

[163] L. von Ahn, M. Blum, N. Hopper, and J. Langford. CAPTCHA: Using hard AI problems for security. In *Proceedings of Eurocrypt, 2003*, 2003.

[164] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, and H. W. Lauw. On ranking controversies in Wikipedia: Models and evaluation. In *Proc. WSDM 2008*, pages 171–182, Palo Alto, CA, USA, 2008. ACM.

[165] J. Wales. Wikimedia and free culture: Past, present and future. In *Wikimania 2006*, 2006. Opening Plenary.

[166] A. Wiggins and K. Crowston. Reclassifying success and tragedy in FLOSS projects. In P. Ågerfalk, C. Boldyreff, J. M. González-Barahona, G. R. Madey, and J. Noll, editors, *Open Source Software: New Horizons*, volume 319, pages 294–307. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[167] A. Wiggins, G. Newman, R. D. Stevenson, and K. Crowston. Mechanisms for data quality and validation in citizen science. In *e-Science Workshops*, pages 14–19. IEEE Computer Society, 2011.

[168] D. M. Wilkinson and B. A. Huberman. Cooperation and quality in Wikipedia. In *Proc. WikiSym 2007*, pages 157–164, Montreal, Quebec, Canada, 2007. ACM.

[169] C. A. Williams, B. Mobasher, and R. Burke. Defending recommender systems: Detection of profile injection attacks. *Service Oriented Computing and Applications*, 1(3):157–170, 2007.

[170] J. Wilson. Volunteering. *Annual Review of Sociology*, 26:215–240, 2000.

[171] T. Wöhner and R. Peters. Assessing the quality of Wikipedia articles with lifecycle based metrics. In *Proc. WikiSym 2009*, New York, NY. ACM.

[172] F. Wu, R. Hoffmann, and D. S. Weld. Information extraction from Wikipedia: Moving down the long tail. In *Proc. KDD 2008*, pages 731–739, Las Vegas, NV, USA, 2008. ACM.

[173] F. Wu and D. S. Weld. Autonomously semantifying Wikipedia. In *Proc. CIKM 2007*, pages 41–50, Lisbon, Portugal, 2007. ACM.

[174] J. Yang, X. Wei, M. S. Ackerman, and L. A. Adamic. Activity lifespan: An analysis of user survival patterns in online knowledge sharing communities. In *Proc. ICWSM 2010*, Washington DC, USA. AAAI.

[175] N. Yee. Maps of digital desires: Exploring the topography of gender and play in online games. In Y. B. Kafai, C. Heeter, J. Denner, and J. Y. Sun, editors, *Beyond Barbie® and Mortal Kombat: New Perspectives on Gender and Gaming*, pages 83–96. The MIT Press, Sept. 2008.

[176] H. Yu, C. Shi, M. Kaminsky, P. B. Gibbons, and F. Xiao. DSybil: Optimal Sybil-Resistance for recommendation systems. In *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, pages 283–298. IEEE Computer Society, 2009.

[177] K. Zickuhr and L. Rainie. Wikipedia, past and present. `http://www.pewinternet.org/Reports/2011/Wikipedia.aspx`, January 2011. Accessed March 19, 2011.