Systems for Collective Human Curation of Online Discussion

by

Amy Xian Zhang

B.S., Rutgers University, New Brunswick (2011)
M.Phil., University of Cambridge (2012)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

Author . . . .
Department of Electrical Engineering and Computer Science
August 30, 2019

Certified by. . . . .
David R. Karger
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Systems for Collective Human Curation of Online Discussion

by

Amy Xian Zhang

Submitted to the Department of Electrical Engineering and Computer Science
on August 30, 2019, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The internet was supposed to democratize discussion, allowing people from all walks of life to communicate with each other at scale. However, this vision has not been fully realized—instead online discourse seems to be getting worse, as people are increasingly drowning in discussion, with much of it unwanted or unpleasant. In this thesis, I present new systems that empower discussion participants to work collectively to bring order to discussions through a range of curation tools that superimpose richer metadata structure on top of standard discussion formats. These systems enable the following new capabilities: 1) recursive summarization of threaded forums using Wikum, 2) teamsourced tagging and summarization of group chat using Tilda, 3) fine-grained customization of email delivery within mailing lists using Murmur, and 4) friendsourced moderation of messages against online harassment using Squadbox.

In a world of abundant discussion and mass capabilities for amplification, the curation of a social space becomes as equally essential as content creation in defining the nature of that space. By putting more powerful techniques for curation in the hands of everyday people, I envision a future where end users are empowered to actively co-create every aspect of their online discussion environments, bringing in their nuanced and contextual insights.

Thesis Supervisor: David R. Karger
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

This work would not have been possible without the support of so many wonderful people. Thank you for your time, your energy, and your care. I dedicate this thesis to you:

- my husband Johnny Hu, who is my biggest supporter and my best friend;
- my parents Randy and Susan and my sister Ally, who I can always count on to be there for me;
- my advisor David Karger, who unfailingly gives solid advice and has been the funnest and most creative co-conspirator;
- my dissertation committee members, who I can also now count on as lifelong mentors—Mark Ackerman, Arvind Satyanarayan, and Ethan Zuckerman—thank you for assisting and encouraging me through this process;
- Mor Naaman, who encouraged me to pursue undergraduate research and then to apply to graduate school;
- Scott Counts, Ed Chi, Jilin Chen, Lichan Hong, Bryan Culbertson, Praveen Paritosh, and Justin Cranshaw, who took a chance on me as an intern and who I can still count on for advice and mentorship today;
- my collaborators in the Haystack Group: Lea Verou, Tarfah Alrashed, Ted Benson, Anant Bhardwaj, Soya Park, Luke Murray, Farnaz Jahanbakhsh, Jumana Almahmoud, and others who helped with feedback and testing;
- my mentees Kaitlin Mahar and Jane Im, who I have to especially thank for your invaluable contributions to this thesis;
- my other mentees Jessica Wang, Janet Sung, Jenny Fan, Sunny Tian, Prateek Kukreja, Joshua Blum, Oliver Dunkley, and others, who patiently put up with me as a newbie research mentor;

- my collaborators outside MIT during my Ph.D., including Chris Schilling and Jonathan Morgan of the Wikimedia Foundation, Connie Moon Sehat, An Xiao Mina, Jenny 8. Lee, Ed Bice, Sandro Hawke, and many others at the Credibility Coalition, Krzysztof Gajos at Harvard University, and Marc Facciotti at UCDavis;

- my team at the Collective Intelligence for Democracy Conference, including Abhishek Srivastava, Daniel Curto-Milet, Pablo Aragon, Berenice Zambrano, Ramla Ayadi, Julio Reyes, and Alejandra Monroy—thank you for believing in Wikum;

- friends and mentors in the broader HCI community at MIT, including Stefanie Mueller and the HCIE group, the VIS group, Rob Miller and the rest of the Usable Programming (formerly UID) Group with whom I spent the first half of my Ph.D., including Juho Kim, Carrie Cai, Elena Glassman, and Philip Guo, and the Center for Civic Media at the MIT Media Lab;

- friends at the Berkman Klein Center for Internet & Society, who expanded my outlook during this last year of my Ph.D.;

- Miranda, Daniel, Joseph, Valerie, and Laszlo, who kindly agreed to be guinea pigs for several of my tools and experiments;

- everyone at Cabot House at Harvard, where I called home for four wonderful years;

- the Google Ph.D. Fellowship and the National Science Foundation Graduate Research Fellowship, which funded my Ph.D. research and study;

- and finally, to all the users who participated in my user studies and people who sat down to be interviewed by me—thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Discussions systems such as email, forums, and chat have been pervasive on the internet since its inception. They contain a diversity of rich information and experiences, including differing opinions on issues, anecdotes, humor, explanations, coordination, and deliberation. Over the course of thousands of comments, even open mathematics problems can be solved [53] and contentious deliberations on Wikipedia settled [154].

As more and more people connect with each other through the widespread adoption of the internet, success stories such as these—where groups of people come together through dialogue to share ideas and achieve more than they could possibly do alone—should be common. Indeed, the internet has revolutionized our capabilities for collaboration at scale in other areas: today it is trivial to read up on a myriad of arcane topics or use one of many customized software libraries thanks to the efforts of thousands of volunteer writers on Wikipedia [361] and open source software developers. And yet when it comes to discussions, the greater scale of participation afforded by the internet has not led to improvements—instead, many would argue that problems with online discussions have only worsened over the last several decades. People are still drowning in information, with few mechanisms for managing or synthesizing large volumes of discourse. A significant proportion of this discourse is unwanted or downright harmful, with clashing norms leading to back-and-forth bickering, and people getting harassed into silence.

In this thesis, I argue that the problems that discussions face online can be traced back to the failure of online discussion systems to *empower users to curate content* as the amount of content has grown. Facing user dissatisfaction and public outcry, large social platforms have outsourced curation to automation and teams of moderators. But these attempts to counter problems have created new problems of their own, from the spread of clickbait and misinformation to censorship of marginalized voices. Instead, what we need is a re-imagination of online discussion systems that places *users in charge* of tending to their shared environments and shared data.

This dissertation serves to illuminate ways that online discussion systems could be redesigned to promote *collective discussion curation* by users. I examine systems for collectively distilling discussion *artifacts* as well as systems for curating message *workflows*. Together, the systems I describe in this thesis demonstrate how we can combat the challenges of online discussion systems in ways that respect human insight and user control. Left unchecked, the problems facing online discussion systems may have far-reaching harmful effects on our society. Thankfully, as a society we are now much more aware of the negative consequences of poorly curated online discourse. With a better grasp of what's been going wrong and why progress has been impeded, we are now ready for something new.

1.1 What is Wrong with Online Discussion Today?

The problems that online discussion systems face today have been festering for decades. Meanwhile, the consequences for ignoring them have grown, as participation in online discussion platforms continues to increase [77] and more forums for discourse, from classrooms [388], to workplaces, to "public squares" discussing news and shaping local and national politics [379], migrate online.

One of these problems is the **scale of discussion content**. The term "email overload" was first coined in 1996 [357], and people were designing systems to combat information overload in online discussion systems such as bulletin boards as early as 1985 [139]. And yet problems of email overload are still around [90, 118] and may have even gotten worse, despite decades of awareness and proposed solutions. A similar story can be told about discussion forums and newer social media platforms, where a single discussion can routinely gain hundreds to thousands of comments. For example, on the discussion board Reddit [277], a quick glance at the homepage shows a post 8 hours old that has already racked up 25,500 replies. Instead of organizing all this content, most systems simply push it down as it gets old, hiding the volume from users by adding page after page. Others just delete old discussions, assuming that users find no value in them.

A separate issue that was also present in the early days of the internet is the prevalence of **unwanted or harmful communication**. Administrators of mailing lists and early forums talked about issues such as "flaming", "trolls", and harassment [55] in the 1990s, as well as email scams and chain letters spreading misinformation [172]. These problems have only escalated in recent years. Due to the migration of more and more of our social lives onto the internet, online harassers have greater ability to inflict harm, and harassment recipients may not have the freedom to opt out of being online [215]. We have also now seen the weaponization of online misinformation spread via social media by state actors and others to sow discord and influence elections [351]. Public opinion has soured towards social media, and many now blame our systems for online discourse for hurting democracy [89] and contributing to rising polarization [257].

It didn't always appear to be this bleak. After the wider adoption of the internet came "Web 2.0" [70], or the emergence of a more social web, which exploded the amount of discourse happening online starting in the late 1990s [302]. This time period was relatively optimistic about the spread of online discourse [16]. Success stories such as Wikipedia, open source software projects, and other forms of "commons-based peer production" heralded more participatory forms of creation [18] as well as innovations arising from web-enabled collective intelligence [216]. Following the early events of the Arab Spring, there was a sense of great potential for newer social media platforms to mobilize collective action [337] and advance democracy [365].

However, the aforementioned problems never really went away, despite various efforts, some of which may have introduced new issues. Faced with growing numbers of users, social media platforms turned to automation as well as teams of paid and volunteer moderators to manage scale and remove objectionable content. However, the many negative externalities of operating these strategies at such vast scale, from perpetuating bias within opaque decision-making processes [82], to not respecting user needs or understanding local contexts, to being easily circumvented by bad actors [147], became glaringly clear as the platforms faced new waves of controversies. Events such as Gamergate [223] and the exposure of Russian state-sponsored "sockpuppetry" online during the 2016 election [292] have contributed to more mainstream awareness [264] of the extent of the problems with social media platforms and highlighted the shortcomings of relying on these stopgap solutions.

1.2 Why Isn't Online Discussion Getting Better?

Why has online discussion as a whole seen so little progress even as our abilities to work together at scale in other areas online have flourished? To answer this question, we can examine the *systems* that we use for conducting online discourse, their major affordances, and how they have evolved or stagnated over the years. What stands out when we look at the major online discussion systems in use today is a *failure to empower users to curate content* as user populations and discussion quantity has grown. This is in stark contrast to systems for collaboration that have successfully scaled up through providing users greater editorial power.

Compared to the power of tools for publishing on social platforms, there are relatively **few tools for user-led curation**, or tools for everyday users to understand, organize, or moderate the large amounts of discussion they see. *Curation* as a practice, originating from museums and libraries and now increasingly applied to digital artifacts, involves selecting found objects, adding new value to them through the process of collecting and annotating metadata [50], and placing them in new contexts and dialogue with other objects [168].

When we consider the social platforms that operate at scale today, many of them focused in their early years on growing their user base as fast as possible and lowering barriers to participation. It didn't matter so much the quality of contributions so long as overall engagement was growing. As competition led to a plethora of places to converse, sites distinguished themselves through their tools for authoring and spreading content. Today, users have unprecedented power to speak and share freely to large audiences online using sophisticated publishing tools. However, as content becomes abundant, the power to curate that content becomes paramount.

In addition, what curatorial tools exist are **limited in expressiveness** and **only available to privileged users**. Administrators and moderators have some curatorial tools at their disposal, namely moderating content or banning users on a case-by-case basis on platforms like Facebook and Twitter. Developers at many social platforms also have great power when designing automation to do the work of sorting or removing unwanted content, such as spam filters within Gmail. Yet these tools fall far short of the range of possibilities that could be encompassed by discussion *curation*, which involves much more than what is generally considered moderation, or keeping out the bad. Discussion curation includes anything from distilling conversations, to signaling norms, to organizing knowledge, to channeling and nurturing pro-social participation [119]—all activities that we do to build positive social spaces and have productive conversations offline [138, 375].

Finally, the concentration of curatorial power into the hands of a select few, either moderators or developers, calls into question the appropriateness and legitimacy of their decisions as a community grows in diversity and size. When the people calling the shots do not understand the context of their decisions and are not even a part of the community they oversee, the result is oftentimes the silencing of already-marginalized populations [107]. What curation tools are afforded to end users by some systems are again limited—primarily forms of flagging [54], voting, and distributed moderation [192].

In contrast, when we look at more successful stories of large-scale online collaboration—systems such as Wikipedia [13] or StackOverflow [8]—these systems have developed sophisticated tools [124], processes [33], and structures [348] to help with maintenance, community building, and organization of content as it grows, as opposed to simply improving tools for authoring more content. In addition, in many of these systems, curatorial powers are available to a much broader swath of the population, and the norms of participation are centered around collective caretaking of a space. Collective curation allows for the scaling of content understanding to occur alongside the scaling of content publishing. As one example, StackOverflow [313] has processes for people to edit each other's questions and answers towards the goal of making them better. This practice is rare in online discussion systems, where individual comments are seen as unalterable except by the author. As a result of these systems' superior capabilities for collective curation, social media platforms such as Youtube and Facebook are increasingly relying on Wikipedia as a way to provide context when false information is posted [97], and StackOverflow has overtaken most mailing lists and forums as the dominant way to get technical information [217, 340].

1.3 From Tools for Publishing to Collectively Curated Social Spaces

Recognizing the lack of evolution in our online discussion systems towards discussion curation, I offer design considerations for how we might address the longstanding problems of online discussion systems. Instead of belatedly trying to fix problems after they are already too big, systems could scale gracefully at the outset by opening up the possibilities that users have to curate online discussion, both for themselves and for each other. Distributed forms of curation can scale as participation scales, as long as there are effective structures for user expression that allow them to be efficiently combined with each other. By empowering end users, they also bring their nuanced and contextual insights towards curation instead of relying on top-down directives.

Framed as an opportunity instead of an automatable chore, curation of a social space transforms users of a publishing tool into co-creators of their discussion space [252]. In addition, instead of treating discussions as throwaway, ephemeral content that can at best languish in disorganized archives, we can recognize that discussion artifacts themselves oftentimes have value worthy of curation to both newcomers and re-visitors.

This involves both a normative shift in a community's culture as well as a shift in technology design. When perceived as citizens of a shared space as opposed to customers of a service [205], members can and may even be expected to contribute to common curatorial tasks, such as documentation [104], norm regulation, and emotional labor [142, 230]—work that is often undervalued or rendered invisible [320, 321]. Much like the automation of content creation itself, automation of curation removes much of what is essential about participating in an online community [107]. By elevating curation to an equal footing alongside discussion content authoring, we can recognize curation's indispensable and uniquely *human* contribution to the discourse production of a community.

Today, we are much more aware of how our decades-old discussion system designs are inadequate for the scale of today's discourse. We realize that the power to speak is ineffective without the power of curation once everyone has a megaphone. We are also more aware of the current limits of and new issues caused by automation and the drawback of putting all of our curation needs into the hands of a centralized body. The time is ripe to develop new systems for online discourse that explore more powerful forms of collective human curation. We urgently need these tools to combat problems that have arisen, from information overload to harassment, in a way that values human input and user control. And beyond fixing problems, we still have yet to realize some of the original aspirations of the social web towards harnessing online discourse to improve public society.

1.4 Systems for Collective Curation of Discussion

In this dissertation, I explore the design space of collective discussion curation by designing and building a series of new online discussion systems that re-imagine outdated discussion designs to give users new curation powers. Within my design of these discussion systems, I consider a number of questions, including 1) What additional curatorial abilities would users like to have within discussion systems?; 2) How would users like to represent and interact with artifacts created from collective curation?; and 3) How can systems structure and motivate user contributions so that they are both easy for users to input yet expressive enough to suit users' curation needs?

The systems I develop explore a series of different collective curation abilities. These include systems that seek to curate *discussion information and presentation*, such as collaborative summarization of threaded discourse (Wikum) and teamsourced markup of chat logs (Tilda). I also examine curation of *message flows and delivery*, such as distributed fine-grained delivery customization of mailing lists (Murmur), and friendsourced moderation of email (Squadbox). These systems were developed after conducting needfinding studies by interviewing and surveying discussion system users about the problems they faced. After the development of each system, I conducted lab experiments and field studies to examine how the systems are used by people in real settings.

1.4.1 Collaborative Summarization of Threaded Forums

Much of the world's factual information is readily accessible today, whether by reading a condensed summary on Wikipedia, accessing an open knowledge base such as Wikidata or Freebase, or quick triaging via a search query. However, there is still a wealth of experiential, contextual, and opinionated information embedded in first-person accounts, advice, and back-and-forth conversations. Unfortunately, this valuable information is often lost within long discussion threads, where the act of sifting through the discourse to get an overview of what was said can be taxing and overwhelming.

The first contribution of my research is a system called Wikum [385] that gives users the power to *summarize* large discussions. Because discussions are often too large for any one person to distill, the techniques I develop, while applicable for individual use, allow summarization to scale with the size of discussion by enabling collaboration.

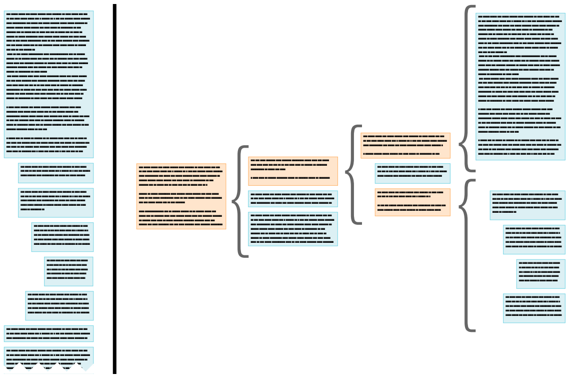


Figure 1-1: **Left:** Discussions are often long and difficult to get an overview. **Right:** Recursive summaries can be constructed to enable progressive hierarchical exploration.

Wikum focuses on the domain of asynchronous threaded discussion, common within email and many forums and comment sections such as Reddit [277] or Disqus [72]. The tool scaffolds the complex task of summarizing and organizing large and unwieldy threaded discussions. Wikum instantiates a crowdsourcing technique called *recursive summarization*, where users build summaries of small sections of the discussion, small sets of those summaries are then aggregated and summarized, and so on until the entire discussion is summarized. This allows the work to be distributed so that each user need only read and summarize a small portion of discussion. Wikum also incorporates techniques from visualization and machine learning to aid users, such as a directly-manipulable tree visualization of the discussion, clustering and tagging suggestions to find related comments to group, and automatic summarization algorithms to assist with summary writing. As shown in Figure 1-1, the result of the workflow is an explorable *summary tree artifact* that allows users to navigate from from a high-level wiki summary to more focused summaries of parts of the discussion to the original back-and-forth forum discussion.

1.4.2 Teamsourced Markup and Note-taking in Group Chat

Struggling to read long conversations can also occur in the case of more real-time conversation such as group chat, popularized in the workplace with tools like Slack, where catching up on missed conversations can be a common occurrence. Through

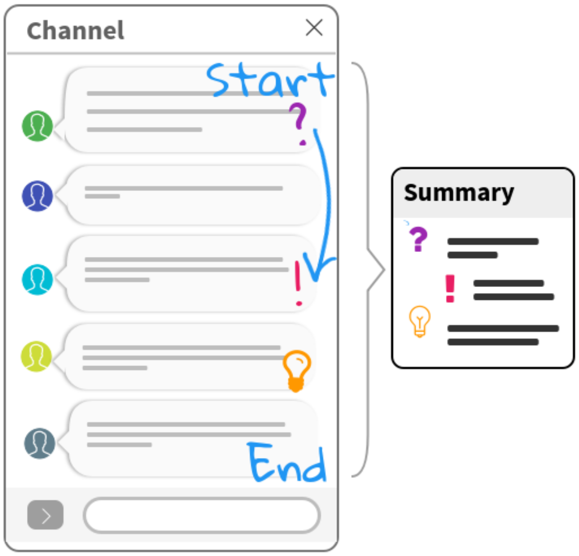


Figure 1-2: Chat messages have little differentiation. Tilda allows users to add markup over chat to enrich the representation of messages and generate summaries for users to use to catch up and dive in points of interest.

interviews with people who use group chat for work, I learned that scrolling is the dominant strategy for catching up, and that making sense of what was said is difficult for users due to the lack of information signals or structure to differentiate chat messages. I then built *Tilda* [381], a tool that provides affordances for rich *markup* over chat with information pertaining to the structure, role, and importance of messages. As shown in Figure 1-2, examples include adding major discourse acts, such as "question" and "answer", linking from one message to another, and delineating separate conversations. Because much conversational context is lost after a conversation is over, Tilda builds in lightweight techniques for *in situ markup* integrated within the chat application, including both text commands in the chat dialog box and direct manipulation via emoji reactions. The markup is then used to automatically construct short summaries of conversations that allow new readers to quickly get an overview and dive in to the original chat messages that are of interest.

1.4.3 Fine-Grained Control over Delivery in Mailing Lists

Not only is it difficult to glean information from long discussions, it is also difficult for users to tell systems what content they want and how they want it. Unfortunately, a significant proportion of online interaction today is unwanted, distracting, untrustworthy, unpleasant, or downright harmful. Sometimes these messages are simply a nuisance, with back-and-forth, repetitive bickering leading to rising incivility or irrelevant messages clogging one's inbox or feed. Other times, they can deeply disrupt someone's life, in the case of online harassment.

Unfortunately, online discussion systems today often do not give users fine-grained control over the mechanisms of content delivery, including the ability to carefully tune what types of messages they receive or how to manage their own attention. On the sender side, users may feel guilty about spamming their recipients but have little ability to target their messages, instead having to settle for everyone getting the message immediately or not at all.

Nowhere is this more clear than in the humble mailing list, a system that has existed for decades but has seen little change [328, 213]. From studies of both workplace and social mailing list communities, I found that, paradoxically, people often wanted more substantive discussion but were themselves too shy to post for fear of spamming. I also uncovered tensions between members due to conflicting ideas about appropriate behavior, partially influenced by how they configured their mail delivery.

Motivated by this work, I developed *Murmur* [377], a re-imagination of the mailing list system that allows members to more finely configure what messages get delivered—for instance, by following threads, individuals, or topics of interest. Conversely, receivers can block topics or only get the initial message of threads, while senders can target to a specific audience or slow a message's propagation. By providing a way for both senders and receivers to fine-tune their delivery, messages can collectively go to only those who want to receive them.

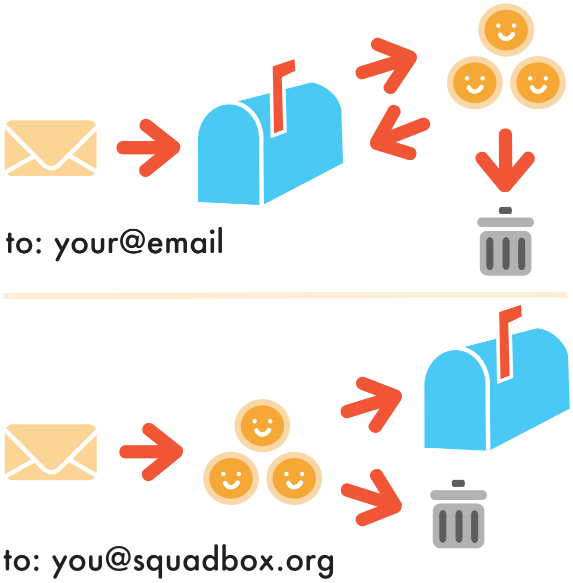


Figure 1-3: Ways to use Squadbox: **1)** auto-forward certain messages from one's inbox to friends, or **2)** create a public-facing moderated account.

1.4.4 Friendsourced Moderation of Email to Combat Harassment

There are cases where it is not enough to provide tools for individuals to manage their own message flow. For instance, in the case of online harassment, where people with an intent to harm flood a recipient's inbox with hurtful or disruptive messages, targeted individuals may become overwhelmed and emotionally vulnerable working alone against determined harassers. When individuals cannot handle moderation on their own, one possible solution is for people to turn towards *networked moderation* strategies, where they can rely on the help of trusted entities or their own community for support and assistance. From interviews with people who face online harassment, we determined that the most personally effective strategy that people use to combat their harassment, besides deleting their account and disappearing off the internet, is to get help from a friend.

From this finding, we then considered how systems can be designed to support *friendsourced moderation*, where a recipient of harassment can forward suspicious messages to friends who then moderate them according to the recipient's wishes. I led a Masters student in the development of *Squadbox* [215], a tool that facilitates friendsourced moderation to combat harassment within email (Figure 1-3). Unlike other unwanted content like spam, harassment is defined in many ways, and recipients have differing preferences on how to deal with harassment. As a result, Squadbox is also designed to be fully customizable by the recipient.

1.5 Thesis Contributions

The focus of my dissertation is on how to design online discussion systems so that end users, collectively but also individually, can have more power to curate their experiences and information within these systems. In order to build systems that actually meet users' needs, I contribute research around three aspects of design across the different systems and domains that I explore.

1.5.1 Empirical Understanding of Desired Discussion Structures and Signals

Discussion systems need ways for more users to access more powerful forms of curation and be able to wield them collectively. One way would be for tools to be built with richer data models for discussion that allow them to contain more context. However it is unclear exactly what information could be better represented in discussion systems to meet user needs. I conduct a number of empirical studies to understand the needs that users have in their online social environments and what metadata and signals, when embedded into systems, could help users realize their needs. Since users have different needs depending on different circumstances, I consider specific needs according to a number of domains and tasks.

For instance, in the case of work communication within group chat, I find that users *visit old content regularly* and want the ability to differentiate chat messages using major *discourse acts*, such as question-and-answer pairs or decisions made after deliberation. I also look at needs within mailing lists communities, finding a desire from *senders* for more fine-grained customization, a feature underdeveloped in many systems. I also separately consider the needs of people dealing with harassment, finding that a *diversity of customizable strategies*, from blocking to forwarding to alerting, were needed to handle different types of messages.

1.5.2 Novel Discussion Presentations and Interactions

I also conduct empirical studies to understand how users ideally would like their discussions presented and how they would like to interact with any curation artifacts, before designing and building novel presentations. For example, I conduct interviews to understand how users would like to view summaries of discussions, finding interest in *dynamic hierarchical presentations*. In the case of Murmur and Squadbox, I consider how users would like to receive messages, in terms of timing, location in their inbox or elsewhere, and presentation of metadata and text. In both the Wikum and Tilda tools, I then design new interfaces for *interactive summaries* that maintain provenance through *hyper-linking* and that express structure through *spatial* relationships between levels of summarization [219]. These presentations allow readers to more easily get an overview of a discussion but then also dive in to read deeper summaries or read the underlying original conversation. In Wikum, this is achieved through an interactive tool with a directly-manipulable *summary tree* artifact.

1.5.3 Techniques for User Expression and Motivations for Curating

Given understanding of what users would like to curate and in what format their contributions should be represented, an open question is how users can best express that information, as enforcing rigid structures can increase hurdles for users [300]. From this, I consider how users can add structure *incrementally* [301], how to reduce the amount of effort required for user input, and how to combine people's efforts so that work can be distributed. In addition, I find in user tests with Wikum that people are oftentimes *afraid to edit each other's work*. Thus, my designs support adding structure so that it is *superimposed* over original discussion artifacts instead of editing or destroying them.

For instance, in the Tilda use case, I focus on techniques to collectively tag conversation in lightweight ways within chat so that participants can do this work while chatting in situ. Similarly, Murmur is designed so that users can customize the delivery of mailing list emails while in their email client of choice using replies. In the Wikum tool, I examine how to break down summarization of a large discussion into smaller tasks where users can build upon and overlay partial summaries on each other's partial summaries using *recursive summarization*. In Squadbox, I explore tools to make moderation easier; for instance, users can build up whitelists and blacklists to automate aspects of their moderation over time.

I also explore potential motivations for users when interviewing people and deploying tools in the wild with different communities. A deep dive into people on Wikipedia who already do a great deal of discussion curation revealed aspects of their work that they would like improved. For instance, when we studied how Wikipedia editors might benefit from Wikum, we found that editors were primarily motivated to use Wikum to reduce their own cognitive load. In Squadbox, we saw that people dealing with harassment had friends who were motivated to help but did not have an easy way of doing so. In Tilda, employees who had coworkers in a different timezone were motivated to keep notes to help their coworkers catch up. In the future, more work is needed to understand motivations over time using long term field studies.

1.6 Thesis Overview

Chapter 2 positions this thesis in the context of related research into the evolution of online discussion systems, the theories we use to understand online collaboration and motivations to contribute to online communities, as well as new interventions, tools, and techniques towards improving collaboration, curation, and discussion.

From there, Chapters 3–6 describe systems and studies exploring novel designs for collectively curating online discussion. These chapters include results from formative needfinding studies, descriptions of system specifications and implementation details, and results from lab evaluations and deployments to communities.

- Chapter 3 introduces the Wikum tool for summarizing large discussion threads. I describe Wikum's recursive summarization process for breaking down the work, and the Wikum interface for exploring summary trees. Then, I present a case study examining the work of Wikipedia editors who must synthesize large deliberations on the platform. Finally, I describe results from a deployment of Wikum to Wikipedia editors.

- Chapter 4 examines group chat systems and introduces the Tilda tool, combining tagging and summarization techniques towards the goal of enriching chat representations and helping users get an overview. I conduct a formative study to understand user frustrations with chat and what presentations and signals are important for catching up. I present results of two lab studies and a field deployment of Tilda with 4 active Slack groups.

- Chapter 5 explores how users can finely control what and how messages get delivered, both as senders and receivers. I describe formative interviews and surveys of mailing list communities, finding tensions due to competing norms. I then introduce Murmur, a re-imagination of the mailing list that allows users to tailor how messages get sent and received.

- Chapter 6 introduces friendsourced moderation to combat the case of online harassment, where individuals are inundated with hateful messages and get help from friends. I introduce the tool Squadbox and also draw from interviews with people who deal with online harassment.

Chapter 7 summarizes design lessons from the four discussion systems and their deployments, and discusses how these findings fit into broader frameworks of curation tools. I also discuss what needs to change in order for collective curation to be adopted more widely. I conclude in Chapter 8 by reviewing the contributions of this dissertation and proposing future research directions.

Chapter 2

Background and Related Work

Systems for conducting discussion online have been around for decades, even longer than the internet has existed. Research into systems for online discussion have existed for almost as long. In this chapter, I cover prior research charting the evolution of our online tools for discussion, the theories and empirical studies that make up our understanding of how people collaborate online and their needs when it comes to discussion tools, and finally novel systems and techniques that could help support collaboration and online discussion.

2.1 Evolution of Online Discussion Systems and Their Lingering Problems

While many aspects of social life online have changed since the early internet days of email, BBSes, Usenet, MUDs, and IRC, what's surprising is how much has actually stayed the same. The dominant method of communication, then and now, is still email. Additional systems, such as IRC and mailing lists, still have relatively broad usage, even as competitors have gained prominence. Why are some of these systems still around? In the cases where systems have evolved, what has been the effects of those changes, and what problems do they still have?

Today, our online discussion tools can be broadly encompassed by the categories of

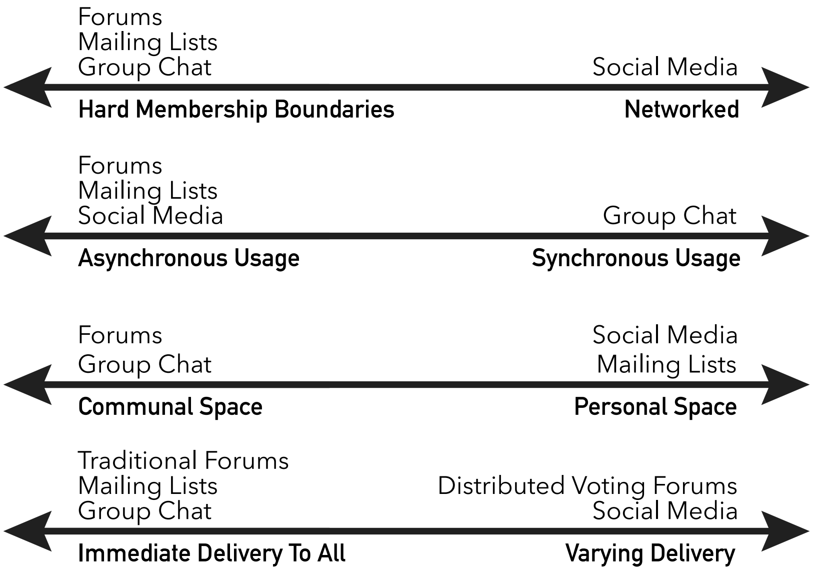


Figure 2-1: The online discussion systems of mailing lists, forums, chat, and social media along four different design dimensions.

mailing lists, forums, group chat, and social media. From their early days until now, each of these systems struggle with the twin problems of information overload and the presence of unwanted content. When we consider each system and the ways in which it is unique or similar to other systems (see Figure 2-1 for several dimensions), we can begin to understand whether and where these two problems become exacerbated.

2.1.1 Email is Still Email

In the 50 years since email was invented, it has become a ubiquitous tool for both private and group communication [23, 213]. Just four years after the invention of email, the first mailing list, MsgGroup, was created in 1971 to help Arpanet users discuss the idea of using Arpanet for discussion. In the 1990s to early 2000s, there was a great deal of excitement over the potential of mailing lists to connect geographically dispersed people in scholarly and professional circles [152]. Studies found that lists allowed highly affective interpersonal interactions [229], encouraged reflection [152], and extended users' social capital [228].

However, even then there were problems, such as complaints about flaming, lurkers, off-topic threads, and information overload [328]. There was also frustration with the need for time-consuming administrative moderation to maintain quality discourse [55]. Some issues with mailing lists in that period simply reflected general problems of email overload [57, 357]. Given the inflexible design of mailing lists, users had no recourse except to unsubscribe when they felt overloaded [328]. Problems were magnified when the messages were deemed nonessential or served a different purpose than regular email, as was often the case for mailing lists [282]. This suggests mailing lists may have exacerbated email overload.

Much has changed in the world since these studies but mailing lists remain ubiquitous despite changing little over the years. Today, there are alternatives to using mailing lists. Google Groups [109] can be used as a mailing list but also offers a web forum experience. Some social media sites have specific affordances for groups, such as Facebook Groups, Reddit, or the now discontinued Google+, with newer features. So why do some people still prefer mailing lists? As seen in Figure 2-1, mailing lists share a great deal of overlap with forums but differ in how they are accessed. Forums exist in a shared space and are accessed the same way by all members, while email users access mailing lists via their personal email client. This allows them to customize how they would like to view and receive messages. Social media systems also offer a personal space to view content via a personalized newsfeed; however, unlike email, social media algorithms cause content to appear out of order or sometimes not at all. Indeed, we find via an interview and survey study (described in Chapter 7) that some people still prefer mailing lists over forums and social media due to characteristics like their greater customizability and greater likelihood for emails to be read.

We also find that problems with mailing lists have continued to plague users. One of the downsides of accessing content via a personal space is that users can develop competing norms about acceptable behavior because of diverging delivery specifications. We find that this leads to tensions between mailing list users. Meanwhile, while email clients are customizable, mailing lists are not—once an email is sent, everyone gets it. Instead, more fine-grained sender affordances could allow users to target emails more carefully and reduce the overall amount of unwanted emails received.

2.1.2 Group Chat: From IRC to Slack

Along with email, chat was one of the earliest forms of computer-mediated communication, and still remains one of the primary ways people communicate online. The first group chat was developed at University of Illinois in the early 1960s to connect users of an instructional system [367]. Chat was initially a popular channel for open source software developers using Internet Relay Chat (IRC). Studies have examined for instance how open source software developers coordinate in distributed teams using chat [299]. Since then, group chat, and its close relative instant messaging, have amassed billions of users world-wide [45, 48].

Researchers have studied the impact of chat systems, including ways that chat can foster intimacy among friends and family [148, 363] and how social norms form in online chat communities [278]. Studies have also shown that chat can lead to unintended consequences, such as a reduction in face-to-face communication, and increased interruption and distraction [35, 101, 56, 156]. Despite bringing people closer and creating a greater sense of community, chat can create artificial distances between people [274].

In more recent years, chat has started to gain adoption in the workplace and more enterprise settings [136, 274, 126]. Tools like Slack, Hipchat, and Microsoft Teams have become popular in the workspace and have opened up an ecosystem of chatbot extensions for connecting chat to other services [186, 368]. As more work is conducted remotely and more businesses move to use group chat, group chat systems have become increasingly important towards improving workplace productivity. However, beyond surface level changes to the appearance of chat clients, much of the underlying structure and functionality of group chat systems have not changed considerably.

Group chat's main difference from other online discussion systems is the greater expectation of synchronous usage rather than asynchronous usage. Synchronous usage can exacerbate problems with information overload when users fall behind on conversations and then have difficulty catching up due to long and messy chat logs. We find that needing to catch up on chat conversations is common as well as cumbersome for users (Chapter 4). Group chat systems also have problems with users receiving irrelevant or unwanted content much like forums or mailing lists due to the inability to customize their space or customize delivery.

2.1.3 A Plethora of Web Forum Systems

Web forums also have a long history, growing from origins in bulletin board systems (BBSes) and newsgroups, first developed in the 1970s and 80s. In the 2000s, popular software packages such as phpBB and vBulletin were developed, allowing anyone with a webserver to easily set up and host an internet forum [74]. These packages included features like allowing flat versus threaded discussion and light moderation and administration tools. Large forum communities on many topics sprung up, each with their own hosting and custom appearance over a generic framework. Many of these packages are still in use today, though their popularity has waned. Similarly, it is possible to attach forums to pages such as blog posts and news articles using systems like Disqus, Wordpress, and Drupal. More recently, new software such as Discourse [71] have updated forum software packages with features common in modern applications, such as infinite scrolling and live updates.

One direction that online forums have gone is towards community Q&A (CQA) sites such as Quora, StackOverflow, or Yahoo! Answers, where contributions are in the form of question-and-answer discourse types. Today, many CQA websites, especially for technical support communities, have overtaken mailing lists and discussion forums as a place for knowledge sharing [340]. These systems incorporate features including tagging, collaborative editing, and marking of solutions to help the community to curate the information available. However it is unclear how well these systems perform for contentious and subjective issues or discourse types other than Q&A. Other platforms for forum discourse are more geared around sharing links to content, include sites like Reddit, Slashdot, or Digg. Unlike traditional forums but similarly to CQA sites, most of these sites incorporate a distributed voting process that alters the placement and visibility of content on the page [192].

As seen in Figure 2-1, while forums like phpBB immediately post content in chronological order, distributed voting forums like Reddit reorder and sometimes hide posts, similarly to social media. While voting is now a common feature in many systems towards reducing information overload, there are documented problems, including underprovision [106], negative feedback loops [40], and harassment campaigns [62]. Social moderation still surfaces only “popular” points to the community. In response, systems could build personal spaces for users to customize what they see but this may also lead to clashing norms, much like in mailing lists, or “filter bubbles” where users self-select into silos containing only one point-of-view [257]. Finally, voting systems have led to issues with harassment and other forms of organized deviant behavior. For instance, within Reddit, harassment tactics such as “brigading” have emerged, where a group of users invade another community to tamper with votes [62]. This is partly because while Reddit has hard membership boundaries for subreddits, users can still vote and post to public subreddits when they are not members.

2.1.4 User Publishing and the Rise of Social Media Platforms

Finally, early applications for sharing user-generated content grew from online bulletin board systems (BBSes) and newsgroups such as Usenet in the 1970s and 80s. This led to an explosion of weblogs on platforms such as Xanga and Blogger by the 90s thanks to the proliferation of end-user web publishing tools. Today, many blogs are still around, with some popular blogs blurring the boundaries between blogs and professional news sites. However, the majority of user-generated content is now hosted on one of a handful of giant social media platforms or “social networking sites” [29], from Facebook to Twitter to Instagram. The scale of participation on these social platforms is unprecedented, so problems can have far-reaching consequences when experienced on one of these platforms.

Studies on social media have looked at what content users share and why [78, 163], and what they choose to not share and why [308, 345, 193]. Research has also looked into the motivations for participation specifically in online groups [279]. Research on Facebook and Twitter suggests that it is used both for information sharing [311] as well as for socialization [259] and self-presentation [239].

While today's social media platforms have some of the same characteristics—and as a consequence some of the same problems—as mailing lists, forums, and chat, one major difference is the lack of hard boundaries regarding communities and membership [79], as seen in Figure 2-1. Instead, interactions form on top of a social network defined by one-to-one follow relationships. This leads to some unique challenges for social media. Norms do not develop on the platform uniformly, as every user sees a different set of interactions. There is little sense of a shared community with the rest of the user base. Also, many different forms of relationships and subcultures are present in the same overloaded space. This can exacerbate issues with clashing norms when different pockets of users come into contact with each other, otherwise known as *context collapse* [221]. Researchers have found that users often self-censor in order to manage their self-presentation [308] and as a result of navigating their identity in different contexts [345]. Online harassment has also flourished as harassers take advantage of the lack of clear rules and boundaries to conduct *networked harassment* [220] tactics such as "dogpiling" [160] to overwhelm individuals.

In addition, content delivery is mediated by black-box algorithms and users view content in a personal space, so that it is not always obvious who will see a piece of content and in what way they will receive it. While partially alleviating information overload, this exacerbates problems with clashing norms. The use of user engagement as a metric feeding these algorithms has itself given rise to new issues with unwanted content such as clickbait, inauthentic engagement via bot-farms and click-farms, increasing partisan content, and misinformation. Social media platforms have struggled to govern this content under one big umbrella using a combination of paid moderators and detection algorithms—on one hand, being scrutinized for biased, inconsistent, or heavy-handed decisions and on the other hand, for not doing enough to protect users from harm.

2.2 Understanding Why and How We Collaborate and Participate in Online Discourse

Much theoretical and empirical research has focused on understanding how users collaborate and communicate and their technology needs for these tasks. This prior work provides an understanding of how we might address longstanding issues with online discussion and informs our design of tools aimed at improving online discussion through collaborative curation.

A significant portion of relevant research resides in the *computer-supported cooperative work (CSCW)* space, a field that got its start in 1984 [117]. CSCW is one of the earliest research communities to examine the design of computer tools for collaboration and coordination between groups of people, building on theories such as distributed cognition [151] and sociomateriality [251] to describe the evolution of technical tools, artifacts, processes, and people in response to each other. Historically, CSCW focused more on small size workteams making use of technology [120]. For instance, much work within *groupware tools* focuses explicitly on technologies to improve productivity and efficiency while collaborating on *a common field of work* [287], sometimes at a distance in time or space [161].

In more recent years, CSCW has moved away from primarily being about traditional work towards becoming inclusive of "*coordinated action*" in general, or action by two or more actors who are working towards a particular goal [201]. This broadens the scope to nontraditional work, including commons-based peer production [18] by stranger volunteers such as on Wikipedia and open source software projects, crowdsourced work such as on citizen science or civic tech projects, collective action such as hashtag movements on social media, or serious leisure [315] undertaken within online communities of interest. In all these cases, the online discourse that happens is a form of *articulation work* [318], or coordination, planning, and all the other work that is done in order to make the primary work function.

Finally, there is a large portion of discussion online today where the discussion does not serve a separate action-oriented purpose but is instead itself the primary goal. For instance, people may provide answers to questions on Quora [275] or participate in debates on Reddit [277] for reasons including socialization or entertainment. Much of the research in understanding these interactions can be encompassed by the broader fields of *computer-mediated communication (CMC)* research and *social computing* [256], research communities dedicated to technology in support of social activity between people. These fields have a great deal of overlap with CSCW.

In this thesis, I consider tools for online discussion in a number of potential scenarios, including both action-oriented and non-action-oriented discourse. However, whether or not the discussion is an instance of work, the role of *curating* discourse of any kind is always itself a form of work, where the common field of work is the site of online discussion. Whether the topic of discussion is a controversial edit on Wikipedia or comments on Facebook about news, there is labor done towards the goal of making those conversations go well. In acknowledgement of this work, curators have been referred to by researchers and curators themselves as “janitors” [323], “custodians” [107], or “gardeners” [294], people whose job, paid or volunteered, is to clean, organize, document, govern, and otherwise tend to a shared digital communication space. Recognizing that this activity is a form of collaborative work allows for us to see the relevant parallels in research on motivations to participate in online collaboration, the design of collaborative work tools, and the creation of collaborative work artifacts.

2.2.1 Motivation to Curate Online Discourse

An important question then is what motivates people to do the work of curation and how well this work can be effectively distributed in the case of online discussion. Much research within social computing has examined motivations for users to contribute towards authoring content in online communities [193] as opposed to curating content, though many of the lessons learned can be applied generally to many forms of user participation.

At the individual level, a *uses and gratifications* [91] approach suggests motivations such as deriving purposive value or social enhancement might lead someone to go beyond contributing content to curating others' content [66]. At the community or organizational level, theories of *organizational commitment* describe when someone develops an affinity with the identity of an organization as a whole [193]. These motivations may change and deepen over time as users transition in their roles and commitments [270]. One space where this has been studied in detail is within the Wikipedia community, where there are strong cultural norms in favor of curation and users accrue social credit and greater powers for explicitly curatorial as opposed to authoring work [32]. We contribute to this work by conducting a deep analysis specifically of those who curate discussions on Wikipedia.

In a similar vein, both *common identity* and *common bond* theories have been applied to motivations to participate in an online community and could be relevant to discussion curation [279]. For instance, in the Squadbox case, we saw that people would be motivated to moderate because of a strong bond with the harassment recipient but also because of common identity, such as being a harassment recipient themselves or sharing a common targeted identity, for instance being a female journalist, with the person being harassed.

Despite these varied motivations to participate, a consistent finding in the study of online communities is the *power law* distribution of participation: the majority of contributions are made by a minority of users [327, 272]. One relevant theory of why more people don't participate is Hardin's "*tragedy of the commons*" [128] which describes the situation where people, acting in their own self-interest when using a shared resource, excessively exploit or degrade that resource as a result. In the case of online discussion systems, the shared resource being depleted is *attention* [304]. Egregious examples of exploitation of attention include spam and clickbait.

The tragedy of the commons has been successfully counteracted in some cases with bottom-up systems of self-governance. In examining many successful empirical cases, Ostrom devises a series of design principles for self-governance [252], including local enforcement, multiple layers of nested enterprises, and an overall emphasis on growing social norms as opposed to imposing rules [253]. These principles have been explored in large-scale online communities such as decentralized governance in Wikipedia [93].

Similarly, I examine more distributed forms of self-governance in the Murmur and Squadbox tools towards resolving issues of competing norms and combating deviant behavior.

Another relevant theory is the "*social loafing*" model, which postulates that an individual's contribution decreases as the size of the group grows [165]. Research suggests that social loafing is detrimental to overall group cohesion [303]. This relates to the tragedy of the commons in that when a common good can be freely consumed, a lack of maintenance oftentimes results [252]. Related theories include the *free-rider problem* or the *prisoner's dilemma*.

Within research on the design of CSCW tools, Grudin ties these concepts to the difficulty of groupware adoption and suggests counteracting these outcomes by creating tools that have equal payoff for all users, emphasizing both individual and collective benefits, and reducing the work required of all users [121]. Failures may also be due to *payoff interdependence*, where one person's use of an application creates positive externalities for others [212], and a critical mass of participation is necessary for all to receive benefit. Within the systems I develop, these recommendations can be seen in how the systems all permit incremental benefits instead of requiring critical mass to receive payoffs.

Studies on social loafing suggest that it is related to the strength of social ties and the perception of risk [303]. Echoing this work, in our study of Wikipedia editors who frequently resolve disputes, we found a hesitance to get involved in cases where there was a chance of reputational risk (termed "wikipolitics"). Designs that rely on strong social ties thus may have lower rates of social loafing; this is reflected in the design of our Squadbox tool that makes use of close ties to encourage assisting a friend facing harassment.

2.2.2 Designing Effective Collaboration Tools

Researchers describe a number of factors that contribute to success in remote collaboration within workteams. This work has implications for discussion curation work, often done at a distance, particularly in the case of tools like Tilda that involve traditional workteams. Olsen and Olsen emphasize the consequences of collaborating at a distance [249], finding that establishing some level of *common ground*, or knowledge that participants have in common and know they have in common, is paramount [250]. Common ground is constantly negotiated on the fly during discussion but is not always explicitly stated.

This has implications for who curates discussion artifacts. In cases where common ground is already high or can be rapidly established due to a higher bandwidth channel (see Media Richness Theory [59]), curation likely can only meaningfully be conducted by discussion participants themselves. In our research on workteams and their use of group chat, we indeed find that it is difficult for non-participants to summarize concluded chat conversations. However, there are many public forum discussions where it is more difficult to establish common ground due to a large and shifting set of participants and asynchronicity of discussion. In these situations, it may be possible for a non-participant to understand the discourse.

A second criteria is the level of *coupling* in work, or the extent and kind of communication required by the work. For most remote collaborations, it is important that the work is loosely coupled, requiring less frequent or less complicated interactions. In the case of discussion curation work, it would thus be important to design collaborative workflows that can break down the work and require only loose coupling. In the deployment of our systems, we have seen some cases where the coupling can be minimal once norms are established, while in other cases coordination is needed, necessitating a way to have meta-discussions about discussion curation.

Other criteria have become less important over the years, such as collaboration technology readiness [24], due to the prevalence of tools. However, *collaboration readiness* is still a key factor, demonstrating the importance of a shared sense of purpose and shared understanding of goals for any collaborative work task, including discussion curation work.

2.2.3 Discussion Curation Artifacts

Finally, a characterization of discussion curation is incomplete without mentioning the sociomaterial artifacts [251] and processes of standardization and formalization that are created as a part of this collaboration work. CSCW researchers have drawn upon the concept of *boundary objects* [314], or objects that lie at the intersection of different communities of practice and help to coordinate their perspectives. This concept can describe common collaboration artifacts such as forms, repositories, or diagrams. Boundary objects are useful for information reuse [209] and organizational memory [3] in CSCW settings and oftentimes rely on some measure of standardization [199], which is a precursor to structure.

Online discussion tools have varying degrees of formal structure. (While communities can always self-impose structures within free-form text, I refer to structures formed at the system level.) Most systems require little structure beyond threads or rooms of conversation and little metadata involved with each post or thread. Exceptions include community Q&A (CQA) sites [4] or systems for structured argumentation or design rationale [214]. However, these systems support only specific forms of discussion. In contrast, general purpose discussion systems involve little standardization and may even resist standardization since they are the go-to medium when other routine processes fail. The benefits of non-standardization can be seen even in CQA sites where there is often a dedicated space for free-form discussion underneath answers.

As a result, the lack of structure and translational context in the raw output of online discussion systems make them non-ideal candidates for boundary objects on their own. Yet despite the difficulty of making sense of unstructured discourse, users do still go back over discussion logs because of their rich repository of organizational knowledge [125]. Indeed, part of the role of a tool like Tilda is being able to translate and distill raw discussion artifacts to a wider audience or different community of practice. Thus, the artifacts created from the discussion curation tools in this thesis have important aspects of boundary objects. As with many boundary objects, they can go through periods of routinization and re-negotiaton over time [210], allowing greater coordination. As an example, we observed people using summaries in Tilda to call out and translate action items and decisions for others not present in the discussion.

A related concept is *boundary negotiating artifacts* [200], which calls attention to collaboration artifacts that don't just sit at a boundary but are used by individuals or communities to negotiate or push what those boundaries are [210], towards the goal of re-configuring the collaborative work. As discussion curation artifacts are created, routinized, and especially used, they can also serve to "iteratively coordinate perspectives" and "bring disparate communities of practice into alignment" [200]. For instance, summarization of a discussion can be used to iteratively coordinate the different perspectives in a deliberation. The final summary artifact is also a structuring device to establish a hierarchy of importance and narrative. Similarly, we saw negotiation between Squadbox owners and moderators regarding best practices for labeling emails and collaborating on whitelists and blacklists.

This conceptualization helps to inform where discussion curation artifacts sit on a continuum of standardization and highlights the potential benefits of discussion curation for coordinating perspectives. While not all boundary negotiating artifacts must eventually become boundary objects, we can see how some of the artifacts from discussion curation tools could inform standard processes and be used to cross even more disparate communities of practice. For instance, Tilda summaries of group chat discussions in workteams could serve as "first drafts" of more formal status updates to managers. Similarly, documentation of harassing messages within Squadbox could be used towards filing reports to police or to social media platforms.

Not all discussions need discussion curation, just as not all curation artifacts need further formalization. The design of the discussion curation tools in this thesis borrows from Shipman and McCall's concept of *incremental formalization* [301], where users can express information in an informal way, and systems can support users incrementally formalizing that information. For example, users of Wikum can choose to not summarize one portion of a discussion at all but summarize another portion with several layers of nested summaries.

There are many benefits to such an approach. Research has shown that users often avoid having to articulate structures or processes explicitly [300], because of cognitive overhead, tacit knowledge that users do not acknowledge, or unwillingness to commit to a structure prematurely. Meanwhile, problems from CSCW tools often arise due to missing social context but computer systems currently must formally encode social information in order to be able to act upon that information. Ackerman describes the distance between the "flexible, nuanced, and contextualized" social requirements of users and what is feasible within technical systems as the *sociotechnical gap* [1].

Though this gap may never be fully ameliorated in practice, systems that can reduce the need for formal information or that make it easier for users to provide information can help [300]. The first step is determining information that is broadly necessary for a task and encoding them as *first-order approximations* [1]. In preparation for developing the Tilda tool, we took this approach to define major discourse acts that users could use to label their chat based on interviews with heavy chat users. Systems also should be designed so that formalization can be defined incrementally and structures can evolve over time. Marshall and Shipman propose *spatial hypertext*, the use of space and visual cues to express relationships, as one way to provide more exploratory structuring.

Incremental formalization is an important characteristic in the discussion curation tools in this thesis. Given the flexibility of the underlying data and wide range of potential tasks, we chose not to go the direction of forcing all discussions into a particular structure, such as in the case of CQA sites. Instead, users can pick and choose where they would like greater structure enacted. We also incorporate spatial information in a visualization in Wikum to represent hierarchical relationships between nodes. In addition, our systems support incremental formalization through *superimposed structure*, or structure that is overlaid on top of original artifacts as opposed to destroying or altering them. This is because we found that users want to read original discussion artifacts and are also reluctant to directly edit other people's statements. Shipman and McCall similarly caution against destructive formalization due to the loss of information [301].

Another possible approach to narrowing the sociotechnical gap is to allow for inferred structure. While much of the information in question is difficult to infer automatically, several of the discussion curation tools support some level of automatic inference based on characteristics of the text. For example, Wikum supports automatic clustering of comments, and Squadbox includes an automatic harassment classifier. In other work, I examine automatic classification of comments in discussion threads according to their major discourse acts [382].

2.3 Existing Research on Techniques and Tools

Finally, I describe prior research into novel techniques and tools that has informed my work as I design new discussion systems. The relevant research topics span many disparate areas, including research on personal information management, novel crowdsourcing systems and processes, visualization techniques, and automatic discourse analysis. In this section, I introduce these fields and situate my research contributions with respect to what came before.

2.3.1 Personal Information Management

The study of *personal information management (PIM)* involves the strategies, tools, and activities people perform with information in order to get what they need done in everyday life [162]. While the discussion information I examine is socially constructed [251], there still exists a "last mile" of information management that involves users interacting with that information on their own.

Notetaking

A common technique for capturing information is lightweight notetaking [339]. Many tools have been developed to improve notetaking in live meetings and lectures, including tools that enable participants to collaborate with shared notes [280, 64, 164, 194], tools for embedding notetaking within multimedia experiences [44, 43], and tools for leveraging meeting recordings to bootstrap notetaking [105, 235]. Despite these tools, there still are not many systems for notetaking within online discussion outside of email, despite many teams moving away from email systems for communication. In my thesis, I describe techniques for lightweight notetaking capabilities within group chat using the Tilda tool.

Email Management

Finally, there exists a great deal of research on PIM tools for email management [357, 19, 58]. Today, a majority of people's workday is spent within email [118, 357]. There has been substantial research on the organizational and retrieval needs of email users. Email users view email as an information repository [356], where they have different strategies for retrieval [9, 332, 241, 305, 325]. Users' needs include email annotation, reliable structure, prioritizing emails, informative overview, flexible sorting, and efficient search [326].

Besides simply managing their own inboxes, researchers have also built systems for users to manage emails on behalf of others or collaboratively within a group. This includes shared inboxes jointly accessed by a team [236] or the use of paid crowd workers to provide personal email management services [183, 184]. While PIM tools mostly focus on email recipients, there are also important sender affordances. For instance, senders would like to hint to their recipients how to respond [114].

When it comes to tools, most email clients today have some sort of filtering, sorting, and searching mechanism. While most existing filter interfaces are focused on explicit metadata within messages, other ways of classification and sorting of messages have been proposed [326, 75, 245] as well as enacted [76]. For instance, research has found that email users tend to see messages as tasks and have a desire to conceptualize email as a task management tool [17]. However, despite all the advancements in email client software, there has been surprisingly little attention paid to group communication tools within email, namely mailing list software. In my thesis, I examine the needs of mailing list users and reimagine mailing list design in light of those needs.

2.3.2 Crowdsourcing and Collective Intelligence

In addition to curating discourse individually, discussion curation can also be conducted collectively. There exists a long line of work on novel techniques and systems to support collaboration, particularly towards discourse curation, though much of the curation capabilities remain shallow. Much of this work can be labeled under the broad concept of *collective intelligence*, or research concerned with group intelligence arising from cognition, coordination, and collaboration [324]. One more recent and relevant subarea is *crowdsourcing* or *human computation* research on complex information processing systems where humans participate and are configurable as discrete computational elements [347].

Voting and Collaborative Filtering

Much of the classic work in collective intelligence examined different forms of crowd judgment systems, such as the "wisdom of the crowds" enabling more accurate estimates of an ox's weight [98]. Today, this idea can be seen in many discussion systems that incorporate some form of community rating process [192] or community flagging process [54] to sort, filter, and moderate comments, threads, and users. These act as a way to both help keep away undesirable content and surface interesting content. However, researchers have documented problems including underprovision [106] and negative feedback loops [40]. Coordinated activity can lead to attacks such as "vote brigading", by calling on members of a community to all down-vote or submit negative content within another community [190]. Collaborative voting may also surface only popular submissions and push down content that may nevertheless be accurate or provide minority opinions, as users interpret "up-votes" as signals of agreement as opposed to accuracy, quality, or relevance.

More recently, some systems have moved away from everyone seeing the same voting outcome towards *collaborative filtering* techniques, where votes from users with a similar background have greater weight. While this may help users see what is personally interesting to them as opposed to more generally, it may lead to "filter bubbles", where users see only the content that reflects their point-of-view [257]. This in turn could lead users to have a false sense of consensus and develop more partisan stances, though research indicates this effect may currently be modest [92].

Structured Voting and Discourse Systems

The flip side of adding voting to discourse is adding more discourse to voting. Some opinion aggregation systems incorporate notions of commentary along with structured judgments. For instance, OpinionSpace asks users a series of rating questions about a topic along with space to leave a comment; comments can then be viewed over a 2-D space of all opinions [83]. Similarly, ConsiderIt allows users to mix and match pro-con lists, placing them on a scale with other users [188].

Other systems have stronger notions of back-and-forth discourse but encode them in a highly structured space. For instance, structured community Q&A (CQA) sites [4] require all discussion to be in a question-and-answer format. There also exists a long line of CSCW systems focused on structured argumentation and design rationale [214]. More recent examples of these include Kialo [171] and the Deliberatorium [181]. While these tools have limited support for free-form discussion, they occupy a space of structured and as a result, aggregate-able, interaction that is richer than voting. However, this limits the kind of dialogue possible and erects barriers to participation. For instance, CQA sites work well for questions that have a clear "best answer" but not as well for questions with many possible answers, such as in the case of opinion-seeking questions or requests for anecdotes [218]. As the focus of this thesis is on general purpose discussion tools, while I incorporate structures including question-answer categories into some of my work, the systems I develop instead permit such structures to arise iteratively based on user input instead of at the outset.

Crowdsourcing Workflows

Finally, one line of work within crowdsourcing has explored how to coordinate crowds of people doing small amounts of work to complete complex informational tasks.

Much of this work has focused on breaking down large tasks into small parts, or *microtasks*, and then providing scaffolding to integrate the parts. Microtask workflows design effective ways to break down complex tasks into manageable, independently executable subtasks that can be distributed to others and executed over time [178, 331]. They have been successfully used for taxonomy-creation [42], writing [333, 334], holding a conversation [197], transcription [196], and scheduling meetings [52].

Researchers have also developed workflows for tasks related to making sense of or synthesizing large collections of data or information, like summarizing books and movies [342], extracting categories and clusters from complex data [10], shortening prose [21], and creating an outline [208] or article [122]. For most of these workflows, the intermediate steps of the workflow are discarded towards producing a final static artifact. In my thesis, Wikum builds on this work by introducing a new workflow for breaking down and combining summarization tasks as well as considering how intermediate work could be externalized in an interface.

2.3.3 Novel Presentations and Interactive Visualizations

Another set of tools and techniques examine novel presentations and interactive visualizations of discussion data. The presentation of information within online discussion systems has changed little over the years. Many online discussions on the web today arrange comments in a linear fashion ordered chronologically. Those that are threaded often use indentation of the comments to indicate reply structure; however this can be difficult to read when there are many replies.

Sensemaking is a process of developing representation and organizing information towards a task, such as decision-making or problem-solving [267]. Building from *information foraging theory*, which posits that people use foraging methods evolved from finding food in the wild to search for information online [266], researchers have worked on imbuing interfaces with cues to improve the *information scent* of content on the page or current path of inquiry [41] for information foragers. As it is common for users to forage through deep discussion threads for information, the concept of information scent is a useful way to think about improving presentations of discussion information.

Alternative Presentations of Discourse

While systems like discussion forums can facilitate collective sensemaking of complex information and multiple perspectives [218], oftentimes this sensemaking happens in spite of the native features within forum software, which make it difficult for users to sample or search through the information space.

Researchers have developed novel alternative presentations to help navigate threads and get an overview of a discussion. For instance, FlashForums provided a thumbnail view of the discussion so users could highlight portions to see the full comments [63]. This sort of thumbnail view provides useful information for readers before they dive into reading a portion of discussion, such as how long a particular thread is or how much back-and-forth there is. Other systems tried mixed-modal visualizations that show threaded conversations in both a tree and sequential way [341]. When it comes to chat, some work focuses on new chat representations, such as allowing people to have threaded conversations in chat [310] or time-based views [96].

Another set of signals that could improve sensemaking capabilities arise from the contributions, navigations, and reading patterns of prior users. This is described as *computational wear*, including *edit wear* to show author interactions and *read wear* to show reader interactions [137], evoking dog-eared pages or well-trodden trails. Techniques such as anchored discussions, as explored in Eyebrowse [378] and NB [388], similarly allow readers to see where on a primary document or on a series of documents the majority of discussion activity is taking place. In addition, techniques such as highlighting or tagging important signals such as emotion within comments [383] can assist with information scent by providing more signposts to readers navigating through and diving in to the discussion space.

Discussion Visualizations

In addition, researchers have explored collecting explicit or implicit signals into more visual representations of discussion. One example is the aforementioned Opinion

Space [83], whose 2-D visual representation encourages exploration of divergent points of view. Another example of a visual tool is Polis[1], a system where users are shown clustered by their level of agreement on a series of user-defined statements.

Researchers have also explored more abstract visual representations of conversations to convey mood, temporal activity, activity by individuals [73], high level content [343], or reply structure [169]. Systems such as ConVisit [146] take the interactivity a step further, allowing users to perform interactive topic modeling over a thumbnail tree view.

Visual representations of discussions can be helpful for sensemaking as they provide an overall picture of the discussion and places to dive in. We incorporate some alternative presentations in our tools, such as a directly-manipulable thumbnail view of threaded discussion within Wikum. One drawback is when visualizations are abstract, such as graph-like diagrams with nodes, they can feel foreign to a certain subset of readers or too complicated for casual readers to go in and manipulate. Another drawback is that large visualizations are difficult to represent in mobile devices. In the future, an alternative approach that could be explored is visualizations that integrate more deeply with text, for instance, sparklines that are at the size of a word [338].

2.3.4 Discourse Analysis and Natural Language Processing

Finally, the *natural language processing (NLP)* research community has a large body of work focusing on automated discourse analysis, text classification, and text summarization, and I incorporate some of these techniques into the tools that I build. While automatic techniques cannot approach human efforts as of yet for many of the curation needs that users have, I consider ways they can *augment* curators' work.

[1] Polis: `pol.is`

Mining Discourse Structure

Some research seeks to mine discussion data or analyze implicit signals in user behavior to extract information about a discussion's structure. Much of this work focuses on the concept of *discourse acts* or *speech acts* that states that each utterance has a major performative function in language and communication. Early work focused only on conversational speech [15, 291]. Since then, researchers have developed standard taxonomies of spoken discourse acts such as DAMSL [317] and DiAML [31]. However, many of these discourse acts for spoken discourse do not translate to online asynchronous mediums. For instance, backchannel responses are not common due to asynchronicity.

When it comes to online discussion, researchers have developed categories for discussions within e-mail [46], online classrooms [87], newsgroups [370], help forums [175], and Reddit forums [382]. Researchers have mined arguments online to learn how people take stances [135]. These kinds of models could be helpful for constructing argumentation or other structures without requiring annotation at the outset of discussion.

In recent years, researchers have become interested in extracting useful information from online discussions. However, many analyses only focus on a particular community [329]. Research in this area has focused on extraction of Q&A content from online forums [47, 144] or characterizing the types and quantity of Q&A content on different community platforms [7, 233]. Other research expands beyond Q&A but still focuses on areas such as technical help forums [175]. In this dissertation, I describe how tools for lightweight tagging in discussion systems could generate data that helps train better models.

Classifying Signals in Discourse

Researchers have built classifiers to detect different kinds of signals in discourse. For instance, many email clients today automatically classify and prioritize emails using machine learning techniques [150, 166, 374].

Other researchers have looked at various communities in order to find patterns of discourse in deliberations. Some have built models for politeness, finding that editors on Wikipedia who are polite achieve higher status through elections [61]. Other research analyzing debate communities such as Reddit's ChangeMyView found that persuasiveness aligned with greater interplay between counterarguments and the initiator [329]. Research on language coordination shows that echoes of linguistic style in responses can determine power differentials [60].

Some researchers have examined classifying harassing, trolling, or otherwise toxic content, using training data created from hand-labeled data [99, 258, 369] or content from existing communities [37]. Researchers have also worked to release data [108] and to better define subtasks within the overall space [167, 354]. However, researchers have qualified this work, warning that such models have documented errors and should not be used without human oversight [6]. Studying existing models, researchers found they could be easily deceived into misclassifying abusive messages [147]. Others found significant differences in data labeling performed by women and men [22], suggesting automated systems can inherit the biases of their data. Finally, researchers suggest that wide differences in norms between communities may make labeled data from one community untransferable to another [22].

Given the criticisms, purely automated approaches to perform activities such as content moderation are not a complete solution in the near-term. Still, there are cases where such models could assist users in their work towards moderation or sensemaking of discourse, such as by suggesting possible tags for a comment or clustering conversations into topics. As another example, I describe later in this dissertation a machine learning model to detect successful resolution of Wikipedia deliberations and assist participants in the discussion.

Automatic Summarization

Finally, there is a long history of natural language processing research on automatic summarization [244]. Some researchers have worked on tools to provide a textual overview or summary of a discussion [276]. Currently, automatic summarization techniques have mostly focused on extractive summarizations [243] which select important sentences from a body of text. This method cannot provide a synthesis of points, such as when paraphrasing multiple redundant comments or determining a resolution from a debate. More recently, researchers have worked on abstractive summarization models [100], which seek to produce novel sentences not present in a body of text. However, most techniques require massive sets of labeled training data [285] which do not exist for summaries of discussions.

Also, most methods are not built for summarizing discussions but instead are for long documents or unconnected user reviews, where more data exists. Of the work on discussion summarization, there includes work on summarizing discussion threads [276, 376], extracting important information from email conversations [286, 371], and analyzing audio conversations [238]. However, automatic summarization still does not perform well enough to be used in practice. Thus, I incorporate it as a potential augmentation that can be ignored by users, such as by highlighting important sentences within the Wikum tool.

2.4 Conclusion

As can be seen, there are many lenses with which to approach online discussion curation. Curation is a form of collaborative work in service to a community, a negotiation towards bringing different perspectives together, and a documentation of knowledge to pass along to others.

This thesis draws upon a rich history of empirical observations of discussion systems over decades of practice, sociotechnical theoretical work drawing insights from sociology and organizational theory, and innovations in techniques and tools as our technological capabilities improve. From these lessons learned, I develop systems that center people in the curation of their conversations and showcase new tools and techniques to help people do this work collaboratively.

Chapter 3

Wikum: Bridging Wikis and Forums towards Summarizing Discussion Threads

3.1 Introduction

Large online discussions involving many participants are pervasive on the web. News and entertainment sites offer comment systems that support discussion of primary content (articles, videos, blog posts) while on other sites the discussion is itself the primary content (Google Groups, forums). These discussions contain a diversity of rich information and may continue to be consulted long after the discussion has died down.

On the downside, such discussions are often "append only." They simply grow, without any kind of organization or summarization. Readers, especially latecomers, need to invest significant time and effort reading to understand a discussion. Though there may be thousands of prior readers, each new reader must individually dig through the same threads of conversation to achieve understanding. There can also be too many tangents and nested layers of discussions to easily navigate. This is so much work that new readers often don't bother, and proceed to post redundant discussion.

Encountering this much information may lead to feelings of overload, due to the unending steam of comments with no narrative or topical cohesion to make sense of their placement or to paint an overall picture. In many cases, the discussion grows so large that it is impossible for an individual to read the entirety of it—why then do interfaces choose to show all of it and with little guidance for exploration?

Current techniques of sorting, filtering, and moderating comments can reduce but not solve these problems. These techniques only select a subset of the comment *texts*; they do not digest or organize their *ideas.* A large number of high quality, popular comments may be upvoted that are all saying much the same thing. Such *redundancy* in discussions may arise independent of quality, making it laborious for participants to identify all facets of the discussion. Similarly, an issue may be argued back and forth and ultimately *resolved*, or incorrect statements may be *refuted.* But these obsolete arguments and incorrect statements remain part of the discussion that a user must wade through to get to the conclusions.

For those seeking a general overview, a short textual *summary* is the traditional solution. But writing a summary of a large discussion will be a massive task, unlikely to appeal to the many readers who do not even bother to *read* the entire discussion. Also, a typical summary offers no way to dive deeper into specific areas based on the reader's interest level or refer back to individual comments.

3.1.1 Contribution

To address these problems, we consider how a *group* of people could individually contribute small amounts of work to refine a large discussion into a *dynamic* textual summary that can be explored at *varying levels of detail.* The main contributions of this work include:

- A new *summary tree* artifact for exploring expandable wiki summaries.
- The *recursive summarization* workflow for breaking down the summarization of a large piece of text.

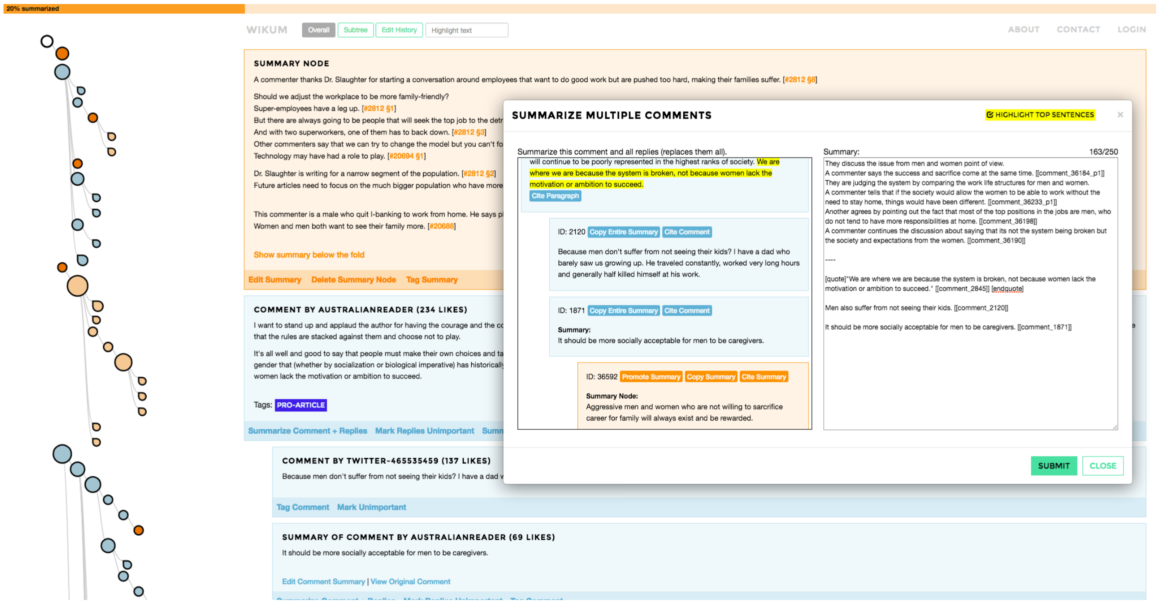
- The Wikum tool for creating and exploring expandable wiki summaries of large discussions.

First, in this work, we present the concept of a *summary tree*, an artifact that is a tree of short summaries of distinct subtopics of a discussion. The summaries are made at different levels of detail so that a higher-level summary covers a greater portion of the discussion. It reflects the paradigm of a good article, where an abstract gives a brief summary of the whole, the introduction summarizes at greater detail, and then individual sections (with their own high level introductions) cover subtopics at even greater detail. By leveraging its online nature, the summary tree is an *expandable* artifact that empowers readers to explore multiple levels of detail, including diving all the way down into original comments. The tree is also akin to topical taxonomies or hierarchical clusterings of items, but in this case each node contains its own substantive information summarizing all nodes nested within.

Second, we design a workflow to create a summary tree using the idea of *recursive summarization* of a discussion, where users build summaries of small sections of the discussion, small sets of those summaries are then aggregated and summarized, and so on until the entire discussion is incorporated into the layered summary tree. Each unit of work requires only writing a short summary of a small number of unsummarized comments or lower-level summaries, so no editor need contribute excessive effort. This way, a group of participants can each do small amounts of work to collectively convert an unwieldy discussion into a short summary of the entire discussion.

Finally, to explore the design space of this process, we developed **Wikum**[1] (a portmanteau of wiki and forum), a system for creating summaries and reading a discussion overlaid with summaries. As seen in 3-1, Wikum combines a directly-manipulatable node-link tree visualization with a view that shows the summaries and comments in focus, as well as a wiki-like editing modal. Readers can explore the discussion, starting at a root summary and drilling into summaries that eventually expand to the original discussion. Editors can edit summaries or contribute additional summaries of unsummarized portions of the discussion.

[1] `http://wikum.org`



SUMMARY NODE

A commenter thanks Dr. Slaughter for starting a conversation around employees that want to do good work but are pushed too hard, making their families suffer. [#2812 §8]

Should we adjust the workplace to be more family-friendly?
Super-employees have a leg up. [#2812 §1]
But there are always going to be people that will seek the top job to the detr
And with two superworkers, one of them has to back down. [#2812 §3]
Other commenters say that we can try to change the model but you can't fo
Technology may have had a role to play. [#20694 §1]

Dr. Slaughter is writing for a narrow segment of the population. [#2812 §2]
Future articles need to focus on the much bigger population who have more

This commenter is a male who quit I-banking to work from home. He says pl
Women and men both want to see their family more. [#20688]

COMMENT BY AUSTRALIANREADER (234 LIKES)

I want to stand up and applaud the author for having the courage and the co
that the rules are stacked against them and choose not to play.

It's all well and good to say that people must make their own choices and ta
gender that (whether by socialization or biological imperative) has historically
women lack the motivation or ambition to succeed.

Tags: PRO-ARTICLE

COMMENT BY TWITTER-465535459 (137 LIKES)

Because men don't suffer from not seeing their kids? I have a dad v

SUMMARY OF COMMENT BY AUSTRALIANREADER (69 LIKES)

It should be more socially acceptable for men to be caregivers.

SUMMARIZE MULTIPLE COMMENTS

Summarize this comment and all replies (replaces them all).

will continue to be poorly represented in the highest ranks of society. We are where we are because the system is broken, not because women lack the motivation or ambition to succeed.

ID: 2120

Because men don't suffer from not seeing their kids? I have a dad who barely saw us growing up. He traveled constantly, worked very long hours and generally half killed himself at his work.

ID: 1871

Summary:
It should be more socially acceptable for men to be caregivers.

ID: 36592

Summary Node:
Aggressive men and women who are not willing to sarcrifice career for family will always exist and be rewarded.

Summary: 163/250

They discuss the issue from men and women point of view.
A commenter says the success and sacrifice come at the same time. [[comment_36184_p1]]
They are judging the system by comparing the work life structures for men and women.
A commenter tells that if the society would allow the women to be able to work without the need to stay home, things would have been different. [[comment_36233_p1]]
Another agrees by pointing out the fact that most of the top positions in the jobs are men, who do not tend to have more responsibilities at home. [[comment_36198]]
A commenter continues the discussion about saying that its not the system being broken but the society and expectations from the women. [[comment_36190]]

[quote]"We are where we are because the system is broken, not because women lack the motivation or ambition to succeed." [[comment_2845]] [endquote]

Men also suffer from not seeing their kids. [[comment_2120]]

It should be more socially acceptable for men to be caregivers. [[comment_1871]]

Figure 3-1: The Wikum interface. Orange nodes are summaries, blue and light orange nodes are original comments. Two of the summaries are expanded, to uncover the comments they are summarizing. An editing window is open to summarize a subthread.

3.1.2 Chapter Overview

In the rest of this chapter, I describe some of the related work around blending discussions and wiki-like editing or summarization and how this work influenced the creation of Wikum. Then I present the major design decisions around the creation of summary trees and recursive summarization, along with details of the implementation of Wikum.

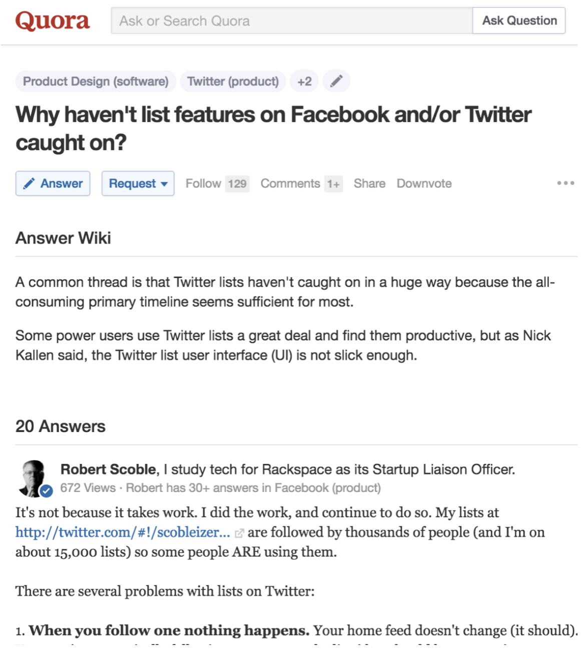
Following that, I describe a lab evaluation to determine the feasibility of our recursive workflow, or how easy it would be to collectively summarize a large discussion using Wikum. Studying the contributions of 20 participants, we found that the same groups of users working in both Wikum and Google Docs were faster at summarizing the discussion in Wikum and also rated it as easier to use. In the Google Doc condition, we saw that users were reluctant to edit other people's work, choosing to append to ever-growing summaries, which ultimately defeated the purpose of summarization. This pitfall was avoided in Wikum as a higher-level summary overlays but does not tamper with other people's work. We performed a second lab evaluation of the created summary trees to understand readers' perceptions of their quality and usefulness. We found evidence from 13 additional participants that Wikum was helpful for quickly getting an overview of the discussion.

I then present a case study involving Wikipedia, where editors must spend long periods of time reading complicated deliberations on Wikipedia talk pages before resolving them. I describe work led by an undergraduate student Jane Im and collaborators Chris Schilling and Jonathan Morgan of the Wikimedia Foundation to understand the problems that editors on Wikipedia grapple with. I then describe results from a field study of Wikum usage by Wikipedia editors.

3.2 Related Work

There are communities and systems that have tried to combine a forum for discussions with a community-maintained wiki or other repository for collecting knowledge [2]. Research on community wikis found that they were useful for managing frequently asked questions [127]. Examples include ExpertNet, a coupled forum and wiki system for government officials to solicit feedback from public experts [246], and Polymath, a successful large scale math collaboration which used a combination of comments, blog posts, and wikis [113]. In Polymath, the two leaders chose to summarize all discussions, a task they found time-consuming but also rewarding. Still, there were issues with newcomers feeling overwhelmed by the discussion. Wikum incorporates some of the design suggestions raised by studies of Polymath [53], including linking from wiki to primary content and citing comments.

Community Q&A (CQA) systems have also experimented with collaborative summarization. For instance, StackExchange permits wiki-like editing of questions and answers [206] and discourages redundant posting. Quora, another CQA system, has experimented with a feature called Answer Wiki (see Figure 3-2) that aims to allow readers to synthesize the answers provided in a Quora question post. However, this wiki box simply sits on top of the answers that appear below with no link between the two. As a result, there is no process or structure for integrating the wiki with the



Why haven't list features on Facebook and/or Twitter caught on?

Answer Wiki

A common thread is that Twitter lists haven't caught on in a huge way because the all-consuming primary timeline seems sufficient for most.

Some power users use Twitter lists a great deal and find them productive, but as Nick Kallen said, the Twitter list user interface (UI) is not slick enough.

20 Answers

Robert Scoble, I study tech for Rackspace as its Startup Liaison Officer.
672 Views · Robert has 30+ answers in Facebook (product)

It's not because it takes work. I did the work, and continue to do so. My lists at http://twitter.com/#!/scobleizer... are followed by thousands of people (and I'm on about 15,000 lists) so some people ARE using them.

There are several problems with lists on Twitter:

1. **When you follow one nothing happens.** Your home feed doesn't change (it should).

Figure 3-2: Quora Answer Wiki feature. An editable wiki text box for summarizing answers sits above the answers.

discussion, navigating from a summary to an original answer, or ensuring the wiki covers the discussion well.

In the other direction are Wikipedia talk pages, where Wikipedia editors deliberate and coordinate their activity on a Wikipedia page [344]. These discussions can be sprawling, with discussions reaching tens of thousands of comments [195]. They are also difficult to make sense of, as there is little support for threading or collapsing of subthreads. Finally, the talk pages have little to no connection to the wiki article they are discussing, for instance to link the outcome of a deliberative discussion to the action made within the wiki.

Some systems similar to Wikum [5, 240] have been proposed that use human work to summarize discussions incrementally. However, none of these systems have had formal user evaluations. Additionally, these systems aim only for a "flat" set of top-level summaries of different topics; unlike Wikum, they do not produce summaries that can expand to reveal different levels of detail to let users drill into specific subtopics. We also *evaluate* our system on both the editing process and the reading experience. Another system explores paraphrasing individual comments within a discussion for the

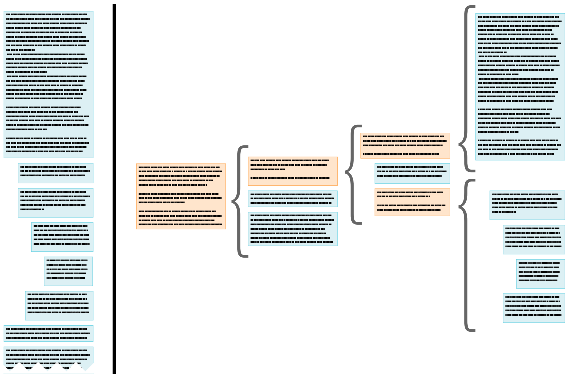


Figure 3-3: From a long threaded discussion (left), we create a summary tree (right). A summary of the entire discussion opens to reveal deeper summaries that open to original comments.

purpose of encouraging reflection [189], but does not have a mechanism for summarizing entire discussions. Deeper reflection can be important benefits of synthesizing conversation, and we are interested in studying how Wikum advances these goals in the future.

3.3 Wikum Design

We begin by outlining the major motivations that informed the design decisions around the summary tree artifact as well as the recursive summarization workflow.

3.3.1 Summary Tree Design

Our artifact and its implementation in the Wikum system aims to combine wikis and forums to address their respective drawbacks. Forums offer no way for someone with little time to get an overview of the discussion, while the condensation required of wikis necessarily drops much of the original detail. To address these complementary drawbacks we could directly combine the two artifacts, as in Quora Answer Wikis, providing a wiki page where a short summary of the entire discussion can be edited. These two components do not connect well though. There is no way to dig down into the summary in order to unpack its origins from the original discussion. The wiki also offers no support for *incremental* summarization—there's no way (aside from reading the entire discussion) to see what has already been summarized and what needs to be added.

We propose a *summary tree* as a more effective bridge that summarizes the discussion forum at *multiple* levels of detail. As shown in Figure 3-3, summaries of small portions of the discussion can be authored, which can then be incorporated into a meta-summary. These meta-summaries can be similarly summarized, until everything is incorporated into a "root" summary of the entire conversation that serves as a starting point for hierarchically exploring the conversation. While other systems have explored creating a "flat" set of summaries of topical portions of a discussion [240], our proposed process of *recursive summarization*, which allows summarization at different depths of the discussion, provides additional benefits. A reader seeking more information can *expand* the root summary into the comments and summaries it summarizes, then choose interesting sub-summaries to expand further. They can dive down as deeply as they like, eventually reaching individual comments. Ideally, the sub-summaries of a summary will cover distinct topics, permitting a reader to focus exploration on topics of interest. As another pathway to accessing "primary source" comments in summaries, our summary tree can include (i) citations to individual comments (and lower level summaries) and (ii) quotes of text from them.

3.3.2 Workflow Design

Making our target artifact a summary tree also suggests a natural approach to constructing it. Starting with the original discussion tree, an editor working alone can choose an appropriately-sized group of related comments to summarize.

Wikum then replaces those comments in the discussion with their summary, treating the summary much like any other comment. Editors can then continue to create new summaries that can distill both previously-written summaries and unsummarized comments, until we are left with a summary of the entire discussion. A reader can reverse this distillation process, expanding interesting summaries to arbitrary depth to acquire more detail.

An important challenge with this process is finding related comments and sum-

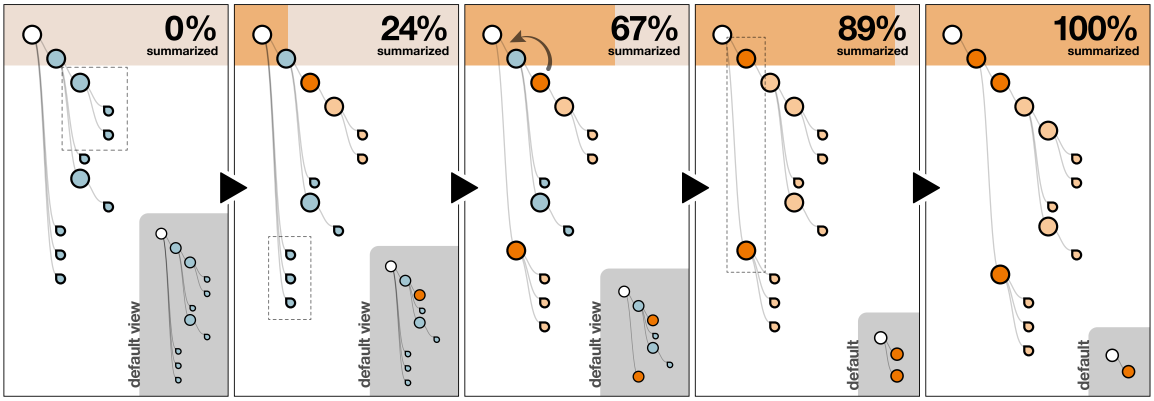


Figure 3-4: Summarization progress for a discussion with 10 comments. Shown here is a fully expanded view of each summary tree state, for illustrative purposes. The bottom right of each panel shows the initial (default) view when the summary tree is in the given state. 1) Initial discussion. 2) Summarizing a comment and its two replies. 3) Grouping & summarizing three root-level comments. 4) Promoting a summary one level up. 5) Summarizing the two root summaries.

maries to bring together and summarize. In the case of threaded discussion, there is a natural grouping heuristic as comments are already organized in a tree structure by reply. Editors can simply pick a small subtree to summarize, where all comments are likely discussing the same topic. Thus, the levels of the reply tree can scaffold the creation of the summary tree. However, even threaded discussions sometimes have comments that have too many replies. Also, given initially threaded comments, the recursive summarization process eventually distills each separate discussion to an individual "root"; these root summaries still need to be gathered and summarized. Likewise, non-threaded discussions have all comments at a single level. To address this, the Wikum system also allow editors to group similar comments at the same level to summarize, using methods like topic clustering, or selecting of adjacent comments (useful for chronological non-threaded discussions).

Even before the summary tree is complete, the summaries that people write in Wikum are *embedded* in the original discussion and contribute towards making the discussion easier to read. In threaded discussions, the summary of a subtree (comment and its replies) lives "between" the comment and its parent. Upon reading the summary, one can expand it to see the comments it summarizes or move on. This can be beneficial to readers because it puts the summaries into context and also provides sensemaking capabilities for exploration of the discussion. Embedding the summaries into the discussion threads also makes it obvious which comments they cover and produces a visual distinction helpful for editors between summarized and unsummarized content.

We designed the summary tree with the goal of supporting effective reading, but our user studies, discussed further, revealed a second benefit. Wikum provides *additive summarization*, augmenting the underlying discussion with summaries. But ability to expand those summaries to reveal the content they summarize, as well as the ability to cite and quote original comments within a summary, makes clear that the material being summarized is *still present*. Thus, the majority of editors' work is *enriching* as opposed to deleting or editing other people's work. This superimposed structure mitigates some of the issues prior research has uncovered around people's reluctance to edit others' work in wiki-like environments [11].

Workflow Efficiency

Recursive summarization permits summarization to be done in small units. But one might worry that the recursive approach significantly increases the overall work requirement as content must be read and summarized at multiple levels. But this is not the case: when each summarization step causes a constant-factor decrease in the amount of as-yet-unsummarized text, the total work done will be little more than that required for one-shot summarization. To see this, suppose that any summary is shorter than the text it is summarizing by a factor of 5. We can therefore conclude that any time an editor reads w words to summarize them, the total text remaining loses $4w/5$ words. If the text starts with W words then it cannot lose more than this before it is fully summarized. Thus, the editors in total will need to read at most $5W/4$ words (of original content or summaries) before the summarization task is complete. And the total number of words written, at $1/5$ of that read by the editors, is only $W/4$. Since comments had to be written once, and are presumably being read many times, the summarization work is proportional to the work users were clearly willing to invest in the discussion in the first place. This suggests that summary tree creation requires only a scalable amount of work.

3.4 Wikum System

The Wikum web interface consists of a tree visualization of the discussion and summaries made so far on the left and a display of selected comments and summaries on the right (see earlier Figure 3-1). Tree nodes are ordered chronologically (within threads when they exist) and can be sorted in other ways. The area of each node corresponds to the length of the corresponding text. Users can select comments by clicking nodes in the tree, which results in the right pane displaying the selected comment and any replies. Users can also select and display disjoint parts of the tree by dragging or Control-clicking. Clicking on a selected node expands or collapses its reply subtree. User-generated summaries are bright orange nodes. Unsummarized comments are displayed as light blue, while summarized comments are light orange to show they have been summarized above. Summaries are collapsed by default and clicking on them reveal the nodes they summarized.

3.4.1 Building the Summary Tree

For readers of a discussion, Wikum lets them see a visual overview, differentiate between summaries and comments, explore into summaries, and jump between conversations. For editors, we provide the same interface with additional affordances for summarization. Wikum enables a number of possible edits to create the summary tree (3-4):

- **Mark as unimportant.** Hides the comment from view. Used for content with no information or interest value.
- **Summarize comment.** Summarizing a longer individual comment is possible. The comment then is replaced with the summary and a link to toggle the original text.

- **Summarize comment & replies.** Summarizes an entire subtree of a threaded discussion into a single summary node. Clicking on the summary node expands it to display the thread subtree.

- **Group & summarize.** Absent threads, we need a way to choose a group of posts to summarize. Even with threading, sometimes a single node may have so many children that it is too much work for one person to summarize. The group & summarize operation lets the editor select a few nodes, then group and summarize them to collapse them down to one node.

- **Promote summary.** If a summary of a subthread has been written, a person writing a summary at a higher level in the discussion thread can promote the lower summary to their position and build on the summary text; this lower summary can be a useful starting point for authoring the higher-level summary.

At the outset, as shown in 3-4, editors may be mostly summarizing a comment and all replies (from a threaded discussion), leaving embedded summaries as signposts to future readers about whether to go down that thread. For non-threaded discussion and later stages of a threaded discussion, grouping and summarizing nodes at the same level that are topically related may be more used.

3.4.2 Creating High Quality Summaries Efficiently

We made additional design decisions to encourage higher quality summary writing. Clicking to summarize one or more comments causes an editing window to pop into view (3-1). This window displays the comment(s) to be summarized on the left, with a text area for the summary on the right.

Important sentence highlighting. We use an automatic extractive summarizer to identify and then highlight important sentences in the content, though this feature can be turned off. This was added to make it easier for people to skim content, though we do not pre-populate the text box with the sentences or allow 1-click transference, due to concerns that it would encourage low quality summaries.

Maximum length restriction. As we noticed people writing lengthy summaries in pilot sessions, Wikum enforces that each summary can be at most 250 words (about half a page) or half the length of the summarized text, whichever is smaller.

Cluster view for comments at the same level. For cases where there are too many adjacent nodes, we provide a clustered view which groups comments that are similar, to help a user select a good group to summarize. This makes it easier to group and summarize topically related comments.

Affordances for citations and quotes. Every node and paragraph within a summarized node can be *cited* in the text summary, which produces a clickable citation when browsing the discussion. Text from original comments can be *quoted* verbatim in the summaries by selecting it and clicking on “Quote". This inserts both the quoted text and a citation to its originating comment. These features were added to encourage summaries that stick to the points made in the discussion. The citations and quotes can also “bubble up" a deeper comment or quote that is interesting or well-written, useful for when readers want to quickly get to high quality comments.

Tag comments and filter by tag. Adding tags to comments is a lightweight task and can also help future summarizers by classifying topics or viewpoints expressed across multiple threads. Comments can also be filtered by specific tags.

3.4.3 System Implementation

The Wikum system is comprised of a front end web interface built using D3, Javascript, HTML, and CSS. It also has a backend component built using the Django web framework and a MySQL database. The homepage of Wikum allows people to paste in URLs to different discussions that kick off a backend ingestion process that adds all the comments to the database. The system currently supports ingesting comments from Disqus, Reddit, and email threads in mbox format. The important sentence highlighting feature was incorporated via sumy[2], a python package implementing the LexRank algorithm for extractive summarization [81]. This algorithm was chosen after experimenting with several unsupervised extractive summarization techniques.

[2] https://pypi.python.org/pypi/sumy

The clustered view for comments at the same level processes the comments and clusters them by first converting each comment into a bag-of-words vector representation that has been TF-IDF normalized. Then the k-means algorithm is used to cluster the vectors. In the cluster view, the cluster with the smallest average distance between pairs of comments is shown first. There is a slider to adjust the size of the cluster, which affects the parameter of number of clusters inputted into k-means.

3.5 Lab Evaluations of Wikum

3.5.1 Study 1: Summarization

We conducted two studies of Wikum to evaluate the process of creating a summary tree as well as the experience reading a summary tree artifact, respectively. In the first study, we sought to understand how long it would take and how easy it would be for a group of people to collectively summarize a large discussion using Wikum versus an alternative system. The second study evaluated the usefulness of the summaries created in the previous stage towards getting an overview as well as people's preferences and strategies around reading discussions using Wikum and our control settings.

In the first study, we evaluated how people summarized content with Wikum compared to more traditional methods to understand the feasibility of the recursive summarization workflow. We recruited 20 participants (mean age 24.9, SD 10.8; 55% female, 45% male) through campus mailing lists and social media and paid $15 for around one hour of their time. All participants reported reading at least one type of online discussion regularly.

Discussion Data

We were interested in seeing how people would summarize content from different discussion topics and types. Thus we selected three different discussions for our study: the comments on an article from the Atlantic called "Why Women Still Can't Have It

All" (Social), a deliberative discussion among members of an academic department about a controversial political event involving their university (Political), and a discussion from the "Explain It Like I'm Five" subreddit seeking to understand a major scientific discovery (Science). Each of these discussions was among the most popular of its category, received many comments from its respective community, and is deeply threaded with many sub-discussions. For the purpose of our study, we pruned the discussions for each condition to roughly equal sizes (removing some of the top level posts and all their replies), aiming for 7,000-8,000 total words or 35-40 minutes of reading given an average reading speed of 200 words per minute [336]. In the end, Social had 84 comments comprising 7,532 total words with the deepest comment 15 levels deep; Political had 67 comments of 7,415 words, with a maximum depth of 14 levels; and Science had 104 comments, 7,375 words, and a maximum depth of 10 levels.

Experiment Design

There were three discussion types, as described earlier, and two system conditions. One system was Wikum, while the control condition was a Google Doc containing the raw discussion text. The text was indented up to 4 levels to indicate threading and then flattened at the 4th level for readability. Google Docs was chosen as a decent approximation to wiki environments. Track changes were turned on to distinguish summaries from original comments so that editors could see each other's work and any text that was deleted by a previous editor. Both conditions included metadata: poster username, number of upvotes, and a unique ID for each comment.

We created three groups and randomly assigned participants to one of them. Each group worked on summarizing two different discussions, one in Wikum and one in the Google Doc, with order counterbalanced. Thus at the end of the study, the three groups produced 3 Wikum summaries and 3 Google Doc summaries, with 2 summaries created per discussion. We chose this experiment design so that we could both compare Wikum versus Google Doc summaries from the same discussion, which controls for that topic of discussion, as well as summaries from the same group, which controls for individual differences in writing ability.

Procedure

User studies were one-on-one, in person, and conducted over a period of two weeks. After completing a short interview and survey about their habits related to online discussions, participants were asked to perform two tasks, limited to 20 minutes each. The goal of each task was to advance the collaborative summarization of one of the two conditions they were assigned, so that at the end, there is a summary of the entire discussion at 250 words or less (half a page). We asked users to work for 20 minutes and no more. Rather than assessing the "natural duration" of an individual's work, we wished to evaluate the *total work* required for summarization, which will likely be distributed among a large number of participants. We kept the time to 20 minutes per task so that each user study would take an hour.

In the Wikum condition, users were first given a 5 minute tutorial on the interface. During the task, we did not give users any particular direction but let them spend their 20 minutes working on what they preferred. In the Google Doc condition, we likewise did not provide directions to users on how to summarize the content. We allowed users to write summaries how and wherever they liked but also encouraged users to be consistent and somehow indicate what was left to summarize to future user study participants. After completing each of the tasks, users filled out surveys on their perceived task load [132]. After both tasks were completed, they filled out a survey comparing the systems and answered some open-ended questions about their experience.

Results

Summaries were completed faster in Wikum than Google Docs by the same group. For each user study condition, we computed the *initial text size*—the number of words in the unsummarized comments plus number of words in the summaries—both at the start of the user task and after its completion. The difference tells us by how many words the user was able to shrink the total amount of initial text. Which

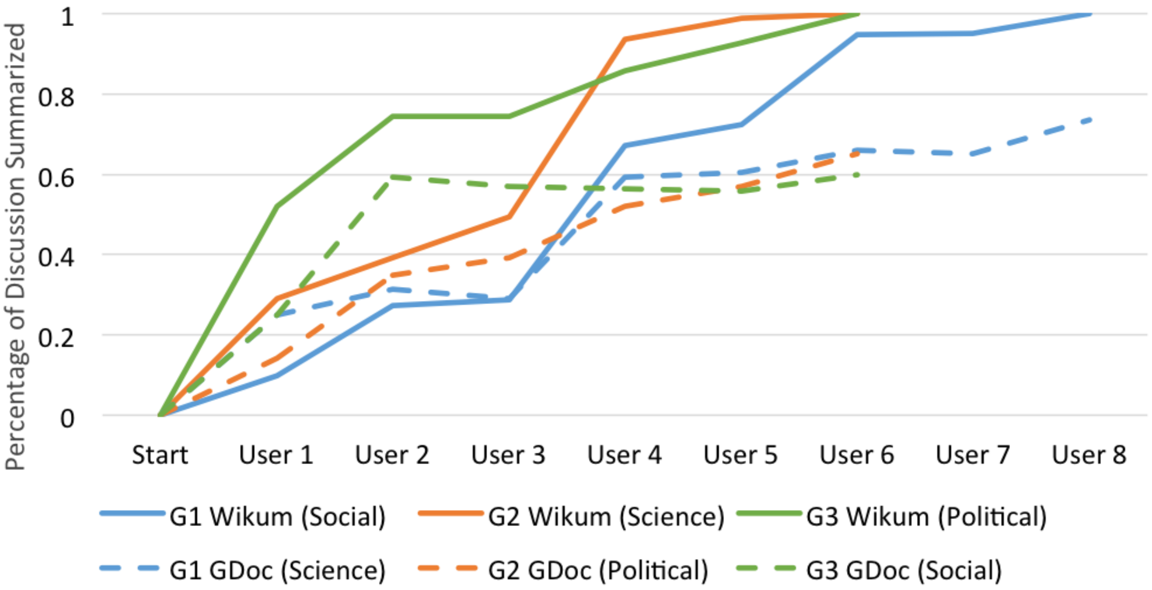


Figure 3-5: Amount of work completed by each successive user in the Summarization stage, by group. Each user amounts to 20 minutes of working time. All Wikum summaries were completed while none of the Google Doc summaries were finished, even with the same group of users editing both.

comments had been summarized was easily defined in Wikum. In the case of Google Docs, we asked users to delineate comments they had summarized in the document, such as using strikeout or marking it "done". We declared a discussion to be fully summarized at the point where the amount of unsummarized content (comments and top-level summaries) totaled 250 words or less. Thus, at the start of our user study, all discussions are at 0% completion, and they reach 100% completion when enough original comments have been summarized so that there are only 250 words to read at the outset.

In Figure 3-5, we show the productivity of the different groups over the course of the study. As can be seen, each group had overall forward progress towards completion in both system conditions but the Wikum condition overall was faster. In total, two Wikum summaries each took a total of 120 minutes, while one took 160, to be completed. The average summarization rate (words reduced per minute) in Wikum was 51.9 while in Google Docs it was 36.3. Thus, in each of the groups, the Wikum summarization of the discussion was completed while the Google Doc summary was still not complete. We chose to stop subjects working on *both* tasks after each Wikum summary was completed because we wanted to use our other user study participants to provide feedback on the Wikum summary qualities as opposed to spending all their study time finishing the Google Docs summaries.

Comparing the 52 word-per-minute Wikum summarization rate with the 200 word-per-minute reading rate we cited earlier shows that summarization is a rapid activity that would demand only a small fraction of the total person-hours devoted to reading a popular discussion.

Users were reluctant to edit others' summaries in both conditions. In the Google Doc condition, 12/20 users chose to only append to an ever-growing single summary that quickly became longer than the 250-word maximum we set. Out of the remaining 8 users, 6 users wrote their summaries interleaved in the comments but did not delete or edit any existing summaries. If users mostly added to summaries and did not delete anything, this would make full summarization impossible since eventually the summary will be larger than the remaining comments. Indeed, as more users participated, we saw overall progress in the Google Doc condition shrink and even plateau in some of the groups, as Figure 3-5 indicates. However, this decline was avoided in Wikum, perhaps because recursive summarization has users summarize other people's summaries *without* destroying their work.

Users spent more time reading in the Google Doc condition. Perhaps as a result of ever-growing summaries in Google Docs, we noticed in the later Google Doc tasks that most users spent almost all the time reading instead of summarizing. As more people edited the document, they spent more time reading the existing summary to determine what was covered, skimming through the comments to find unsummarized content, and figuring out how to incorporate their findings back into the summary. One editor said:

> "*Using the Wikum was so much easier...I knew what people had done...With the Google Doc it was this massive 40 page document. I got lost on what people had summarized and what needed to be summarized.*"

Some editors did not bother to read previous summaries and then accidentally

	Group 1	**Group 2**	**Group 3**
Wikum	1037 (Social)	1310 (Science)	497 (Political)
Google Docs	769 (Science)	1073 (Political)	771 (Social)

Table 3.1: Total number of summary words written by users in Wikum versus Google Docs within each group.

	Social (G1)	**Science (G2)**	**Political (G3)**
Summary Nodes	13	20	6
Citations	25	36	4
Quotes	0	7	0
Tags	6	1	5

Table 3.2: Total number of times each item was used or created in each of the three Wikum summary trees.

summarized portions that had already been summarized. Like Google Docs, wikis also lack this kind of scaffolding for summarization. However, some of these issues might potentially be mitigated with a more defined style guide or set of instructions.

Users overall wrote more summary text in the Wikum condition. Perhaps as a result of needing to spend less time coordinating other people's edits in the Wikum condition, users overall wrote more in the Wikum condition, as can be seen in Table 3.1. Though the amount of time spent and the people were kept constant per group, users overall wrote 2,844 words in Wikum versus 2,613 words in Google Docs. As described in the earlier Workflow Efficiency section, this additional summarization did not add much work compared to the 7-8,000 words in the original discussion. In the case of Group 3, the one group where Wikum users wrote less, the Wikum condition had one early participant who chose to summarize a large subthread in one summary. As readers complained about this in the second study, this suggests that in the future we should only allow editors to summarize limited chunks of discussion at a time.

In the case of summarization, more may not always be better. A thousand words of summary is around two pages long, which may be more than someone is willing to read. However, because of recursive summarization in the Wikum case, users can read a 250-word summary of the entire discussion and drill in to get more detailed summaries.

Earlier editors set the norms for later editors in Wikum. We noticed during the user study that the decisions made by early editors in Wikum, such as to use citations or quotes, set the norms for future editors, echoing prior work on norm setting in communities [173]. This led to different styles of summarization emerging in different groups. For instance, early use of citations and quotes led to more use of these features in the SCIENCE Wikum summaries, while it was not used at all by early POLITICAL editors (Table 3.2). The same was true for the case of adding tags. In the future, this could be more scaffolded, for instance by requiring some number of citations per number of comments being summarized.

The convergence of norms happened to a lesser extent in the Google Doc conditions. For instance, people would use different ways of signaling they finished summarizing a comment in the same document. Some users also chose to write their summary of a particular sub-discussion interleaved among the comments even if others had been contributing to a single summary at the top of the document. Later contributors tended to do this as the single summary got more unwieldy, and unsummarized comments were further from the summary at the top of the page.

Editors made use of the citation and quoting features. Many users chose to add citations in the summaries (Table 3.2). Several users liked the ability to cite, saying:

> "*The way in which you can cite paragraphs and posts is very usefulâĂę to have that kind of chain of custody, like from where does this information come from?*"

However, the quoting feature was used less often, possibly because it was less discoverable, as one needed to drag-and-select text before a "Quote" button showed. In the future, we could add "Quote" buttons next to highlighted sentences. Some editors used quoting and citing as a way to minimize editorializing and deflect lack of understanding of the content:

> "*Obviously someone who has a physics background would be better over me. Me summarizing this comment, I don't know if I would trust me. That's why I tried to quote a lot and really cite what was going on.*"

The same user went on to say:

> "*...People might only read my summary, they might not read the actual comments, so I felt pressure to make sure you've accurately summarized the comment.*"

For her, citing and quoting was also a way to point readers to original content and to also self-check that she was summarizing the comments faithfully.

Users reported that summarizing content they disagreed with took more effort. Some users expressed frustration with comments they disliked, with one editor saying:

> "*What I really wanted to be like was, this comment is stupid because it said this, rather than writing an unbiased thing. I think some of my summaries were a little snarky.*"

A different editor mentioned working harder but also that she was more interested:

> "*It was more interesting to summarize comments that I disagreed with because it requires you to try to understand their point of view as much as possible...I already know my own point of view.*"

Reflection and learning gained from summarizing other people's opinions [189] could be an additional side benefit of Wikum. As in Wikipedia, there may be value in educating editors about maintaining a so-called *Neutral Point of View (NPOV)* during summarization work [224].

Overall feedback on summarization. Users overall felt that the recursive summarization process helped to break the task down to something manageable, with one editor saying:

> "*A lot of times I would look at a comment and all its sub-comments and be like, well I can't summarize all that, it's really overwhelming. But then I was able to drill down into the sub-sub-comments and...get the whole comment [subtree] and sub-comments into my head at the same time, write a summary, and then go a level up.*"

From the post-study survey, users indicated that they preferred conducting summarizing using Wikum over Google Docs (t=3.02, p<0.01). Users also found Wikum easier to use. Survey results related to task load [132] revealed a significant difference when it came to physical demand, with Wikum overall causing lower physical demand (t=2.07, p=0.05, paired t-test). This may be because many users complained about needing to scroll more in the Google Doc condition. Likewise, Wikum showed lower temporal demand (feeling hurried or rushed during the task) (t=3.11, p<0.01), possibly because it look less time to get started editing in Wikum as opposed to Google Docs. Editors in Wikum also self-reported higher performance on the task (t=2.37, p<0.05).

3.5.2 Study 2: Reading and Exploration

In the second part of the user study, our goal was to assess whether a Wikum summary tree is a useful tool for quickly getting an overview of a discussion. We recruited 13 more participants (mean age 28.0, SD 9.7, 72.2% male, 27.8% female) via the same methods described in the previous stage. As before, all participants said they read at least one type of online discussion regularly. Participants were compensated $10 for around 40 minutes of their time.

Experiment Design

Before seeing any summaries of the discussion, the first author of this paper read over the three discussions and extracted a list of main points made in each. Care was taken to include points made throughout the discussion including in sub-threads that were deeply nested. As we only showed editors a subset of the original discussion in

Study 1, the author also looked over the comments that were pruned from the original discussion in order to come up with another list of points that were not in the study, but that could plausibly have been.

We designed a 2-factor user study where each participant was given three tasks, each limited to 20 minutes. For each task, the participant was given one of the three discussions and one of three interface conditions. One condition was the Wikum interface with the embedded summaries that users made in the prior stage. A second condition (DocSummary) was a Google Doc containing the summaries and the original comments also created in the prior stage. Summary text was colored purple, while deleted comments were faded gray. Original comments that had not been processed by the first stage participants were colored black. Summaries were left wherever users placed them in the preceding stage, whether that was at the top of the document or interspersed throughout the discussion. We also provided easier navigation to the different summaries using the Google Docs outline feature. The third condition (NoSummary) was a control, consisting of a Google Doc containing only the raw discussion with no summaries. The assignment of the discussion topics and interface conditions as well as the order was counterbalanced.

Procedure

In each task, the participant was given 10 minutes to try to get an overview of the discussion. During this time, the authors observed how participants chose to explore the discussion in the different interfaces. Then, without the discussion in front of them, they were presented with a list of 12 points, 6 of which had been mentioned in the discussion and 6 of which had not. Participants were not told the number of points that were false. They were asked to select points they remembered being brought up in the discussion. At the end, participants completed a survey about their experience and discussed their experience reading using the different interfaces.

Conditions	Precision	Recall	F1
Wikum	0.90	0.67	0.78
Google Docs Summary	0.88	0.63	0.72
Google Docs No Summary	0.81	0.58	0.65

Table 3.3: The results of Study 2 between the three conditions.

Results

Most explored the Google Doc linearly, while there was a mix of strategies using Wikum. For the NoSummary condition, almost all participants read linearly down the page, with most running out of time before they read even half of the discussion. For the DocSummary condition, most users also read linearly down the page, though some users chose to focus on reading the summaries and skip over or skim the comments that were in gray. Others chose to read original comments, even if they already read the summary.

In the Wikum condition, people had a mix of strategies. Several users (5/13) chose to expand the discussion tree fully and read linearly down the discussion on the right, sometimes scrolling past some subthreads, but overall treating the Wikum interface exactly how they would a Google Doc. Others (4/13) chose a breadth-first approach from the root, reading summaries at each level and only expanding summaries when they deemed it necessary. Some users chose to expand everything at the outset but then focus on the summary nodes using the tree visualization, going from the root to the leaves (3/13) or from the leaves to the root (1/13). Many of the users who focused only on the summaries chose to stop reading well before the 10-minute cutoff, suggesting they had already achieved full comprehension.

Users recalled points made in the discussion more accurately in the Wikum condition. From the recall test, as seen in Table 5.1, Wikum performed slightly better than DocSummary on the measures of precison, accuracy, and F1 score, and both summary conditions performed better than NoSummary. However, none of the differences in scores between the three different conditions yielded a statistically significant difference (with $p<0.05$), likely due to the small sample size and the variations in topic, quality of summary, order of conditions, and different reading strategies and speeds. Thus, these results suggest that summaries are indeed helpful for getting an overview in a short amount of time, and that users were able to get an overview using Wikum as least as well as using Google Docs. Though the difference between Wikum and Google Docs with summaries was not significant, recall that all users were familiar with the Google Docs interface but had only a few minutes to learn the new Wikum interface. One user said:

> "*A big chunk of the time went into understanding the Wikum interface itself - more than half. If I had seen this interface 5 or 10 times I would be familiar with it.*"

Some people preferred reading linearly while others enjoyed drilling in. The Wikum interfaces defaults to hiding comments underneath a summary. Some people disliked needing to click to open up a summary, saying:

> "*[I would like to] have more control about what I was going to read, as well as look at the scrollbar to know the amount of content ahead of me.*"

As a related issue, some people enjoyed the tree visualization, while other people found it overwhelming. While the tree visualization seems a useful feature for editors, it may be less necessary for readers of a summary tree.

People opened summaries to read comments for different reasons. Some people said they would read comments below a summary if it was poorly written or too short because they did not trust it. For instance, one person said:

> "*That's the scientist in me. I need to see, is this comment really saying that? I didn't want the summaries to influence my take.*"

Other times, readers actually thought the summary was well written and thus it piqued their curiosity:

> "*I was more likely to read the individual comments on the good summaries. The summaries went into depth, so I figured there was more discussion there. Good = interesting, so I wanted to learn more.*"

Overall feedback on reading and exploration. When it came to their experience reading and exploring the comments using the different interfaces, users rated Wikum the highest (4.2 on average on a 7-point Likert scale from 0 to 6), with DocSummary second best (3.6), and NoSummary the worst (2.5). The difference between the Wikum and DocSummary was not statistically significant ($p<0.05$) while the difference between those two conditions and the unsummarized one was significant (Wikum: $t=-3.04$, $p<0.005$, DocSummary: $t=-3.05$, $p<0.005$). Users were also asked to grade summary quality on a 7-point Likert scale. Overall users felt the Wikum summaries were of higher quality than the Google Doc ones (4.5 versus 3.5 on average respectively), though this difference was also not statistically significant. Thus our results suggest but do not conclude that Wikum provided benefits for readers over the Google Doc, and affirms that summaries are a useful way for readers to get an overview of a discussion.

From post-study interviews, users mentioned that the Wikum summaries were more succinct while Google Doc summaries went on for too long. This is despite the fact that the total text in all the Wikum summaries was actually greater for those users. One user said of the Wikum summaries:

> "*It felt good on a few comments - it was very noticeable...that there was a large amount of text just swirling around a few simple ideas, and the summary got it simple. Like into a tweet. That was really, really nice. I wish everything could be summarized like that.*"

Another user said:

> "*I felt it was helpful for Wikum but not really in the Google Doc. There, there were people rambling...It was kind of a mess. Because the summaries were right there in Wikum and directly related to the comments, [they were] much smaller summaries and a lot more helpful.*"

A different user echoed that the Wikum summaries were shorter, and complained that the highest-level Wikum summary was too abstract so that he had to dig deeper to understand portions of the discussion (which Wikum is specifically design to support). This could be related to the preference some people had for reading linearly.

3.5.3 Design Implications

During the summarization stage of our user studies, we saw that Google Docs was too underconstrained so there were many opportunities for editors to go astray and set poor norms. However, even though Wikum has more constraints, we realized that some additional scaffolding could guide editors towards creating better summaries while still maintaining Wikum's flexibility. One editor was worried about too much rehashing, saying:

> "*If you encourage a summary every time you have a parent or child, you'll just have crummy summary on top of crummy summary...Trying to encourage only summaries when you have a certain depth or breadth to the tree would go a long way.*"

In the other direction, one user chose to summarize a large portion of the discussion at once, producing a low quality summary. Later readers of this summary tree were surprised to find so many comments under that summary. This indicates that there may be an optimal range of discussion size that should be summarized in a recursive summary. Too small and the recursive summaries feel too incremental and repetitive to a reader. Too big and the summaries have poor coverage and hide a great deal of discussion. Wikum could also suggest groups of comments to target for summarization via heuristics or machine learning. These could include the start of a self-contained subthread, a clear shift in topic or participants, or a discussion devolving into arguing.

Another issue that came up was around the difficulty of summarizing opinionated content, especially content the editors disagreed with. Computation techniques in detecting language that is objective versus subjective [359] or determining opinionated or emotional sentences [360] could be a useful addition to a summary editing box to help editors monitor the language they use.

When it came to the reading experience, many readers in the study talked about trust as an important factor while reading the summaries. If they did not trust that the summaries were accurate or had good coverage, they felt they needed to read more of the original content. Distrust of wikis and other crowd-editable content can sometimes be mitigated with design [179]. This was one reason for our emphasis on citations and quoting. Other ways to improve trust could involve showing information such as number of edits, total time spent writing a summary, number of contributors, or percentage of original discussion cited. We could also introduce a form of social moderation, allowing readers to rate summaries on accuracy.

Finally, our study reveals future areas for experimentation with different presentations of the summary tree. Some readers liked the information that the tree visualization provided but others felt it was overwhelming or too disconnected from the text. Some ideas to explore include trying to integrate information that the tree provides directly into the discussion text, such as toggle controls, breadcrumbs, or even simplified subtree thumbnails. Views were also mixed on the preference for an expandable versus linear reading experience, echoing prior work in the hypertext literature around jumping around using links [112, 295]. Unlike a graph-structured hypertext however, which can pose significant navigation challenges [254], Wikum is likely easier to navigate since it is hierarchical. Additionally, one can ignore the expandable nature of Wikum and pre-expand everything, as we saw a few readers do, and read linearly. In the future, we could make this even easier by allowing readers to set how much of the summaries they wish to have autoexpanded upon load.

3.6 Case Study: Deliberation and Resolution in Wikipedia

In this section, I describe a case study focusing on deliberative discussions within the English Wikipedia community, conducted in partnership with the Wikimedia Foundation and led by Jane Im. I advised this project and participated in conducting the interviews. The Wikipedia community is a relevant community to study regarding discussion summarization as there already exists a process where difficult-to-resolve issues are deliberated at large, and then the oftentimes long discussions are perused, summarized, and resolved by an independent editor. Looking this process over the course of 7 years, we find that nearly a third of discussions never get resolved. Through interviews with frequent summarizers, qualitative content analysis, and machine learning models, we uncover the major problems with Wikipedia's current deliberation resolution process.

3.6.1 Introduction

The study of online processes for deliberation and resolution touch upon many areas, including open democratic initiatives and civic participation [255], as well as virtual teams [158], open source development [198], and online community maintenance [271]. One such area is Wikipedia, a place where almost all conflict is resolved through online deliberation. The stakes for deliberation can be high—for instance, the addition of two paragraphs about a city on its Wikipedia page can lead to significant changes in tourism [140]. As a result, conflicts arise on the platform regularly [180, 372], mirroring conflicts around contested information in the world. Prior research has often focused on "edit wars", or back-and-forth edits on Wikipedia articles, as well as on article talk pages [322], where editors go to informally resolve an issue, as signals of conflict and resolution. However, there are also various formal resolution processes for disputes that cannot be resolved informally, with differing layers of escalation. The study of these formal processes can reveal insights about factors leading to resolution as well as areas of friction, towards the design of better processes and systems for online deliberation and resolution.

To better understand online deliberation, we investigated one of the primary formal processes on English Wikipedia for deliberation and resolution of content and policy disputes—the Request for Comment (RfC) process. Using RfCs, editors who cannot resolve a dispute may publicize their deliberation to the broader Wikipedia community to invite participation, sometimes culminating in a *closing statement* by a neutral editor that summarizes the discussion and makes a resolution.

We created a novel, comprehensive dataset of 7,316 RfCs from English Wikipedia dating from 2011 to 2017, parsed to separate out closing statements, authors, and reply structure. This dataset is released publicly for the research community. [3] We employed a mixed-methods approach by analyzing this data quantitatively as a whole as well as qualitatively by selecting a random subset of 40 RfCs to manually inspect. To inform our analysis, we interviewed 10 of the most frequent RfC closers to understand their motivations and considerations when deciding whether to close an RfC.

From the complementary sources of data, we examined what major factors in the RfC process result in failure to come to a resolution. Not all RfCs require a formal resolution by a closer; instead, some may informally end due to overwhelming agreement by participants or withdrawal of the RfC by the initiator. In our dataset, we found that 57.65% of RfCs end up getting formally closed through the addition of a summary statement resolving the dispute. However, of the 42.35% of RfCs with no formal resolution, we found that 78% had no participant activity to informally end the RfC—in other words, that *a full one third of all RfCs in our dataset were left stale*. A prevalence of stale and unresolved disputes may mean that effort put into discussion is wasted and time is lost waiting for resolution.

From interviews and qualitative analysis of our dataset, we uncovered reasons for why these RfCs do not get formally closed, including factors such as poorly articulated initial statements by inexperienced discussion initiators, lack of interest from third-party experienced Wikipedia editors, and excessive bickering or contentiousness during the discussion.

Using these factors to inform a series of features, we developed a model to predict whether an RfC will go stale based on information about the page before the RfC initiation as well as what transpired over the course of participation in the RfC. When trained and tested on our dataset, the best model achieved 75.3% accuracy, an improvement of 8.1% over a baseline of simply predicting that it will not go stale. We find that the most informative features as to whether an RfC will go stale are the size and shape of the discussion along with features related to interest and expertise level of participants. Furthermore, we consider how well such a model performs as an

[3] `https://figshare.com/articles/rfc_sql/7038575`

RfC progresses in time after its initiation.

3.6.2 Background on Wikipedia Governance and Deliberation

Processes for Resolving Content Disputes

Broadly, there are two types of disputes in Wikipedia, content-related disputes, which include policy disputes, and user conduct disputes, and numerous formal and informal mechanisms for achieving resolutions for each type. While our focus is on content-related disputes, the line between the two types can blur, as user conduct issues can arise in the course of a deliberation about content. When a dispute cannot be resolved by the involved members on their own, there are a number of ways to receive outside help. First, Third Opinion (3O) is reserved for content-related issues between exactly two editors, and is a relatively informal process for getting an outside opinion. In comparison, the Dispute Resolution Noticeboard (DRN) is used for disputes involving more than two parties or when 3O does not resolve the dispute. Volunteer moderators on the noticeboard provide suggestions and mediation towards the dispute, but this process is primarily limited to simple disputes that can be quickly resolved. If the dispute escalates, there is Formal Mediation, which is provided by a panel of experienced mediators called the Mediation Committee (MedCom) who resolve Requests for Mediation (RfM) once they are filed.

At any point in the escalation of dispute resolution processes, editors can turn to Requests for Comments (RfCs) by writing up a proposal or question on the relevant article talk page and then inviting comment by the broader community by posting to various noticeboards. For this work, we chose to focus on RfCs as it is one of the more common formal processes for resolution due to its flexibility, and because it involves a number of editors across Wikipedia due to the gathering of input from the broader community, as opposed to places like 3O or DRN.

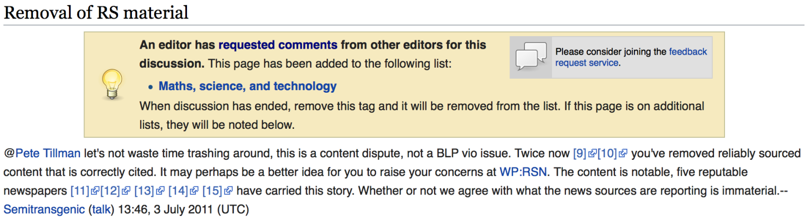
Research on Deliberation in Wikipedia

Researchers have analyzed the deliberative discussions that happen on Wikipedia, finding evidence of both constructive behavior and pitfalls [344, 288]. Conversations on talk pages can create long chains of back-and-forth responses in a format much like threaded forums [195]. Analysis of talk page communication found that it scales up to help manage conflict as the number of editors grow [177]. Qualitative analyses of deliberation on Wikipedia found a high level of analytic discussion focused on problem analysis [25], while other work has found examples of debates around information quality [319]. However, researchers have also found lower levels of social aspects of deliberation such as respect and consideration [25], and other researchers found cases of power plays when policies are unclear and advocate for more tools to support the consensus process [187].

While most existing work focuses on informal coordination and communication, in this work we turn to more formal mechanisms for conflict resolution. There exists some analyses of these formal discussions for the case of Articles for Deletion (AfD) [103], though there the number of participants per discussion is generally small and the emphasis is on voting [330]. There are also both formal and informal processes for managing user roles and promotion within Wikipedia. Some of the formal processes involve deliberation, such as the Request for Adminship (RfA) process for selecting administrators on Wikipedia. Research has shown that a model considering factors like strong edit history can predict which users will be voted in as an administrator [32]. In this work, we shed light on a particular type of informal editor role that has not been studied in detail, which is that of frequent RfC closer.

Wikipedia Tools for Discourse

There have been many efforts to improve the interface of talk pages and build tools for consensus. Some have targeted the unstructured nature of talk pages, which can cause difficulty for newcomers, and have developed lightweight tools to add structure [289]. Others have developed models to predict different dialog acts in Wikipedia [88],



Removal of RS material

An editor has requested comments from other editors for this discussion. This page has been added to the following list:

- **Maths, science, and technology**

When discussion has ended, remove this tag and it will be removed from the list. If this page is on additional lists, they will be noted below.

Please consider joining the feedback request service.

@Pete Tillman let's not waste time trashing around, this is a content dispute, not a BLP vio issue. Twice now [9][10] you've removed reliably sourced content that is correctly cited. It may perhaps be a better idea for you to raise your concerns at WP:RSN. The content is notable, five reputable newspapers [11][12] [13] [14] [15] have carried this story. Whether or not we agree with what the news sources are reporting is immaterial.--Semitransgenic (talk) 13:46, 3 July 2011 (UTC)

Figure 3-6: Screenshot of an RfC started by using the RfC template tag `{{rfc}}`.

which could also lend greater structure. Within the MediaWiki platform, interfaces have been developed that make talk pages more like question-answering systems or threaded forums, such as Flow[4] and LiquidThreads[5]. Researchers have also sought to support consensus-building on Wikipedia, including tools to summarize behavior and track conflicts as they unfold [187].

3.6.3 Introduction of Requests for Comment

Requests for Comment (RfCs) are a common process use by Wikipedia editors, or volunteers who write Wikipedia articles, for requesting input from uninvolved editors concerning disputes about policy, guideline, or article content. It is a formal way to attract more attention to a problem that is not resolvable with local discussions, and uses a system of centralized noticeboards and bot[6]-delivered invitations to advertise discussions.

Initiation: The process for RfCs starts with a content dispute that has already been discussed in a talk page but has not been resolved. At that point, an editor can start a new section within the talk page. Using the RfC template tag `{{rfc}}`, the initiator writes a neutral statement in the form of a proposal or question outlining the issue at hand, optionally selecting one or more topical categories as well, as shown in Figure 3-6. Any Wikipedia editor can be the *initiator* of an RfC.

Dissemination: After the initiator adds the RfC template tag to the page, a

[4] `https://en.wikipedia.org/wiki/Wikipedia:Flow`

[5] `https://en.wikipedia.org/wiki/Wikipedia:LiquidThreads`

[6] Bots are computer-controlled user accounts that help maintain pages: `https://en.wikipedia.org/wiki/Wikipedia:Bots`

Wikipedia bot called Legobot assigns the RfC an ID and posts the RfC on the RfC list page pertaining to that category. Legobot also notifies a random subset of editors that are watching pages or lists related to the RfC, such as editors who have volunteered via the Feedback Request Service[7]. There are currently 2,360 editors listed as volunteers, though editors also provide a limit on how many notifications to receive a month. Anyone may also post the RfC manually to places such as Village Pump[8] forums, various noticeboards, talk pages of relevant WikiProjects, and talk pages of related articles or policies, in order to invite more discussion from people not already involved.

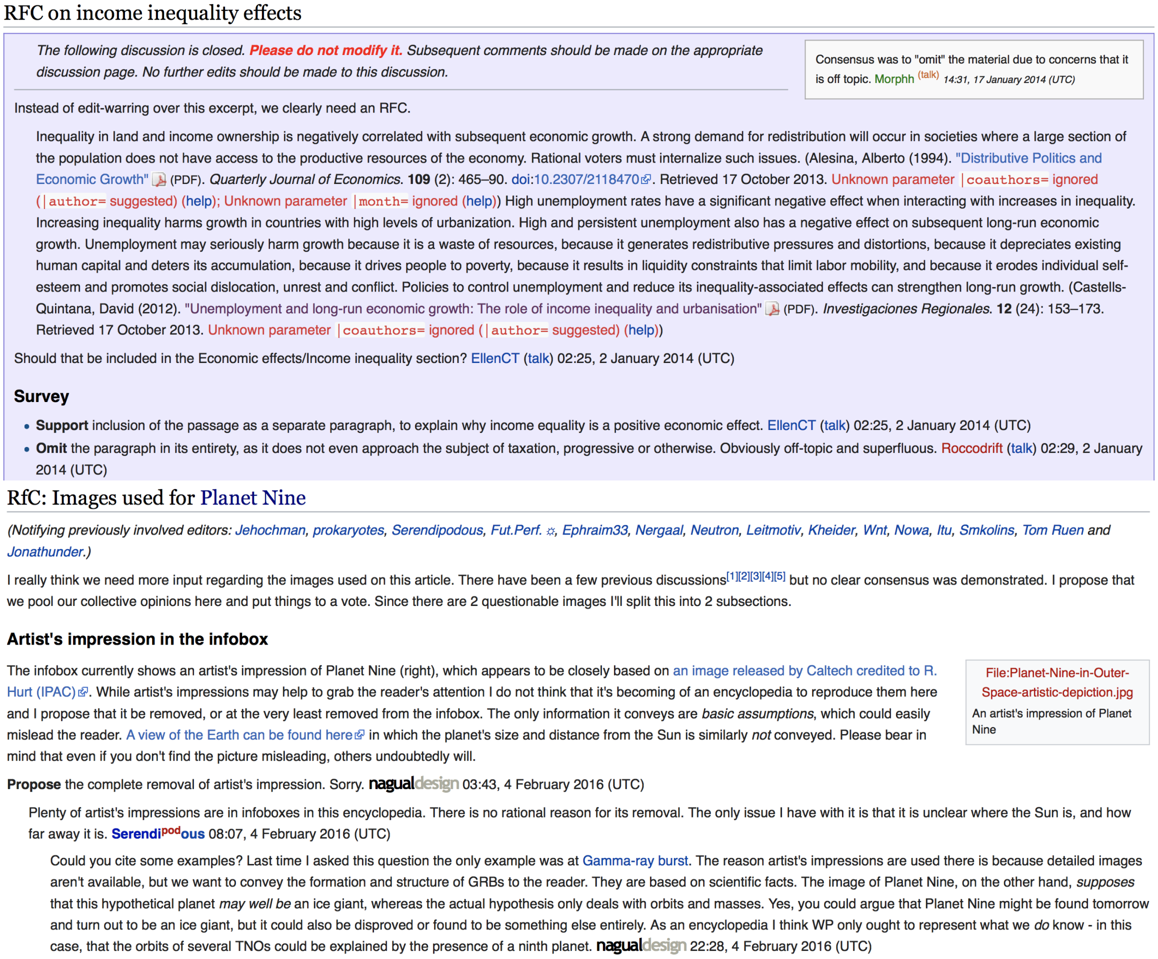
Discussion: Once initiated and publicized, the discussion unfolds in a threaded fashion using indenting. Some RfCs also include a section for users to indicate their position in a polling process. The default length of an RfC is 30 days, after which Legobot automatically removes the RfC template tag, and it gets removed from RfC lists. Participants can delay this removal if discussion is still ongoing or they can revive the RfC by re-adding the tag later. The RfC may be closed early if consensus is clear before 30 days, though a general practice is to wait at least a week for input. Although anyone can participate in an RfC, the system is targeted towards getting input from uninvolved editors who can provide unbiased opinions to help resolve the dispute.

Closure and Conclusion: After a certain period RfCs can conclude with three type of endings, which are a (i) ***formal closure***, an (ii) ***informal end***, or (iii) simply be left ***stale***. These three endings are organized in Table 3.4. (i) *Formal closure* is a general process for relatively more contentious debates, requesting an uninvolved third party to close and mark the end of the discussion. Anyone may post the RfC to the Wikipedia Administrators' Noticeboard/Requests for closure[9], a clearinghouse where frequent closers go to find unclosed RfCs. A closer closes the RfC by adding the templates `{{archivetop}}` and `{{archivebottom}}` along with a closing statement surrounding the RfC as shown in Figure 3-7.

[7] https://en.wikipedia.org/wiki/Wikipedia:Feedback_request_service

[8] https://en.wikipedia.org/wiki/Wikipedia:Village_pump

[9] https://en.wikipedia.org/wiki/Wikipedia:Administrators%27_noticeboard/Requests_for_closure



RFC on income inequality effects

The following discussion is closed. ***Please do not modify it.*** *Subsequent comments should be made on the appropriate discussion page. No further edits should be made to this discussion.*

Consensus was to "omit" the material due to concerns that it is off topic. Morphh (talk) *14:31, 17 January 2014 (UTC)*

Instead of edit-warring over this excerpt, we clearly need an RFC.

Inequality in land and income ownership is negatively correlated with subsequent economic growth. A strong demand for redistribution will occur in societies where a large section of the population does not have access to the productive resources of the economy. Rational voters must internalize such issues. (Alesina, Alberto (1994). "Distributive Politics and Economic Growth" (PDF). *Quarterly Journal of Economics.* **109** (2): 465–90. doi:10.2307/2118470. Retrieved 17 October 2013. Unknown parameter |coauthors= ignored (|author= suggested) (help); Unknown parameter |month= ignored (help)) High unemployment rates have a significant negative effect when interacting with increases in inequality. Increasing inequality harms growth in countries with high levels of urbanization. High and persistent unemployment also has a negative effect on subsequent long-run economic growth. Unemployment may seriously harm growth because it is a waste of resources, because it generates redistributive pressures and distortions, because it depreciates existing human capital and deters its accumulation, because it drives people to poverty, because it results in liquidity constraints that limit labor mobility, and because it erodes individual self-esteem and promotes social dislocation, unrest and conflict. Policies to control unemployment and reduce its inequality-associated effects can strengthen long-run growth. (Castells-Quintana, David (2012). "Unemployment and long-run economic growth: The role of income inequality and urbanisation" (PDF). *Investigaciones Regionales.* **12** (24): 153–173. Retrieved 17 October 2013. Unknown parameter |coauthors= ignored (|author= suggested) (help))

Should that be included in the Economic effects/Income inequality section? EllenCT (talk) 02:25, 2 January 2014 (UTC)

Survey

- **Support** inclusion of the passage as a separate paragraph, to explain why income equality is a positive economic effect. EllenCT (talk) 02:25, 2 January 2014 (UTC)
- **Omit** the paragraph in its entirety, as it does not even approach the subject of taxation, progressive or otherwise. Obviously off-topic and superfluous. Roccodrift (talk) 02:29, 2 January 2014 (UTC)

RfC: Images used for Planet Nine

(Notifying previously involved editors: Jehochman, prokaryotes, Serendipodous, Fut.Perf. ☼, Ephraim33, Nergaal, Neutron, Leitmotiv, Kheider, Wnt, Nowa, Itu, Smkolins, Tom Ruen and Jonathunder.)

I really think we need more input regarding the images used on this article. There have been a few previous discussions[1][2][3][4][5] but no clear consensus was demonstrated. I propose that we pool our collective opinions here and put things to a vote. Since there are 2 questionable images I'll split this into 2 subsections.

Artist's impression in the infobox

File:Planet-Nine-in-Outer-Space-artistic-depiction.jpg
An artist's impression of Planet Nine

The infobox currently shows an artist's impression of Planet Nine (right), which appears to be closely based on an image released by Caltech credited to R. Hurt (IPAC). While artist's impressions may help to grab the reader's attention I do not think that it's becoming of an encyclopedia to reproduce them here and I propose that it be removed, or at the very least removed from the infobox. The only information it conveys are *basic assumptions*, which could easily mislead the reader. A view of the Earth can be found here in which the planet's size and distance from the Sun is similarly *not* conveyed. Please bear in mind that even if you don't find the picture misleading, others undoubtedly will.

Propose the complete removal of artist's impression. Sorry. **nagual**design 03:43, 4 February 2016 (UTC)

Plenty of artist's impressions are in infoboxes in this encyclopedia. There is no rational reason for its removal. The only issue I have with it is that it is unclear where the Sun is, and how far away it is. **Serendipodous** 08:07, 4 February 2016 (UTC)

Could you cite some examples? Last time I asked this question the only example was at Gamma-ray burst. The reason artist's impressions are used there is because detailed images aren't available, but we want to convey the formation and structure of GRBs to the reader. They are based on scientific facts. The image of Planet Nine, on the other hand, *supposes* that this hypothetical planet *may well be* an ice giant, whereas the actual hypothesis only deals with orbits and masses. Yes, you could argue that Planet Nine might be found tomorrow and turn out to be an ice giant, but it could also be disproved or found to be something else entirely. As an encyclopedia I think WP only ought to represent what we *do* know - in this case, that the orbits of several TNOs could be explained by the presence of a ninth planet. **nagual**design 22:28, 4 February 2016 (UTC)

Figure 3-7: Comparison of a formally closed RfC (top) and one that is not (bottom). Formally closed RfCs have a purple box surrounding the thread and a grey closing statement box. On the other hand, RfCs that are not formally closed have no such template.

	(i) Formally closed	**(ii) Informally ended**	**(iii) Stale**
Ended by whom	Uninvolved editor	Participant, initiator, or uninvolved editor	None
RfC tag is removed by whom	Closer	Participant, initiator, or uninvolved editor	Legobot
Closing template exists	Yes	No	No
Dispute is resolved	Yes	Yes	No
No. of RfCs	4,086 (57.65%)	672 (9.48%)	2,329 (32.86%)

Table 3.4: Differentiation of the three possible outcomes of RfCs.

For the remaining RfCs without these templates, there are two possibilities as to what was the outcome of the RfC. First, the RfC could have been (ii) *informally ended* on purpose by participants, the initiator, or another editor by removing the RfC tag manually. This might happen because the initiator reconsiders and chooses to withdraw the RfC, or an obvious consensus may lead participants to agree to withdraw the RfC. Second, the RfC could have (iii) gone *stale*—that is, while waiting for further participation or a formal close, there is a period of no activity for 30 days, and the RfC never gets closed by an individual. In this case, Legobot would remove the RfC tag after 30 days of inactivity, effectively withdrawing the RfC if no one bothers to open it up again. For the rest of this work, we use the term ***"unclosed"*** to describe (iii) where RfCs remained stale, without any kind of closure and ***"closed"*** to describe both (i) and (ii). We use ***"formally closed"*** and ***"informally ended/closed"*** when we want to indicate (i) or (ii) respectively.

Any editor on Wikipedia can be the *closer*, formally or informally, of an RfC close an RfC; however, formal closers tend to be more experienced editors on Wikipedia due to their grasp of Wikipedia policy and greater perceived authority within the community. Also, some RfCs do require closure by an administrator if the close involves action that can only be done by an administrator, such as deleting an article or unprotecting a page.

Post-Close Review: In the case of formal closures, especially for more contentious ones, it is not uncommon for participants to question the close or ask for details. This usually takes place on the closer's user talk page or more rarely the close can be challenged by posting to the Administrator's Noticeboard. There is no specific venue for reviewing RfC closes, unlike AfD decisions[10], so it can be difficult to determine what happened after an RfC ended. Another way to relitigate an RfC is to hold another RfC at a later point in time. While it is frowned upon to hold an RfC soon after a closed RfC on the same topic, they can generally happen since consensus may change over time.

3.6.4 Analyzing Seven Years of RfCs

Data Collection

As there is no archive of links to all past RfCs, to gather as many RfCs as we could, we focused on edits left by Legobot, a bot that is automatically triggered when the RfC template tag `{{rfc}}`[11] is added to a discussion to create an RfC. Using this strategy, we collected a dataset of 7,316 RfCs beginning from 2011, when Legobot began running, to the end of 2017. We used this dataset to analyze characteristics of contributors as well as the lifecycle of RfCs, from initiation to a final outcome. From this dataset, we can determine RfCs that have been (i) *formally closed* using a template as shown in the left of Figure 3-7. Analyzing the dataset and the interviews revealed, however, that among the RfCs that did not have the template, not all were simply left stale. Thus, we differentiated between (ii) *informally ended* RfCs and (iii) *stale* ones by tracking the revision history to find when the RfC tag was removed and then retrieving the user account that removed the tag. If it was removed by Legobot, we considered it stale; if the RfC tag was removed by an editor, it was treated as informally ended. While not perfect—for instance, participants might choose to withdraw their RfC but neglect to remove the RfC tag—this method represents our

[10] https://en.wikipedia.org/wiki/Wikipedia:Deletion_review/Active

[11] https://en.wikipedia.org/wiki/Template:Rfc

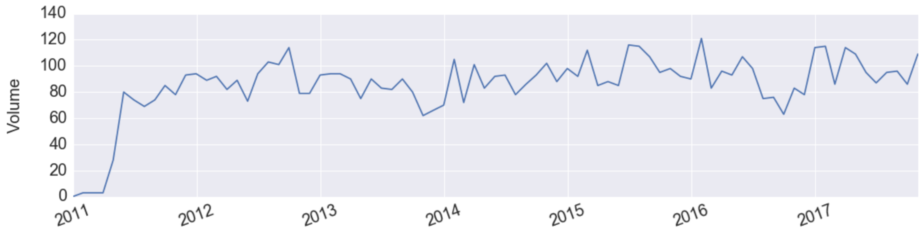


Figure 3-8: The number of RfCs initiated each month in our dataset from 2011 to end of 2017.

best approximation from the data available to reconstruct what happened.

We were able to categorize 7,087 RfCs out of 7,316 RfCs using this method. 57.65% of the RfCs ended up formally closed while 42.35% have no formal resolution. Among the unclosed ones, 78% (2,329, 32.86% of all RfCs) remained stale without any closure, while 22% (672, 9.48% of all RfCs) were informally ended. Among the 672 informally ended RfCs, 522 were ended by participants or initiators who took the tag off while 150 were ended by uninvolved editors. Although the former is considered the norm, inspecting the 150 RfCs showed that in some cases uninvolved editors take the RfC tag off if they believe it is no longer necessary or should not have been created. Since in these 150 cases an editor ended a discussion by taking the action of removing the tag, we counted it as informally ended.

Participation, Participants, Topics, and Dynamics of RfCs Over Time

We characterize our RfC dataset to demonstrate how the RfC process works currently and how it has evolved over time.

Initiation: From looking at Figure 3-8, we can see that the number of RfCs initiated over time has remained fairly steady since mid-2011, with 86.5 initiated per month on average across our dataset. Table 3.6 provides information about the initiator population, which overall is smaller and more experienced than the participant population.

Dissemination: Table 3.5 shows the number of RfCs initiated within each category from 2004 to 2017. These category counts can give us a rough understanding of areas of relatively higher and lower levels of contention within Wikipedia. When

RfC category	**No. RfCs initiated**
Politics, government, & law	2650
History & geography	2573
Biographies	2123
Wikipedia policies & guidelines	1767
Uncategorized	1732
Society, sports, & culture	1634
Art, architecture, literature, & media	1601
Maths, science, & technology	1165
Religion & philosophy	949
Wikipedia style & naming	749
Wikipedia proposals	634
Economy, trade, & companies	585
Wikipedia technical issues & templates	381
Language & linguistics	372
WikiProjects & collaborations	259

Table 3.5: Number of RfCs issued from 2004 to 2017 by categories. One RfC may have multiple categories, for example, `{{rfc|econ|bio}}`.

	Initiators	**Participants**	**Closers**
Total number of people	3,346	14,815	759
Percentage of administrators	7.41%	5.11%	23%
Avg (σ) number of edit counts	23,432.16 (74,417.6)	14,055.43 (56,749.5)	39,759.46 (89,639.2)
Median number of edit counts	4,590.5	1,257	17,556
Avg (σ) account age (days)	3,076.63 (1,338.2)	2,260.05 (1,226.1)	3,289.3 (1340.2)
Median account age (days)	3,230.81	2,331.71	3,635.67

Table 3.6: Overall information about RfC initiators, participants, and closers. The values for initiators and participants was calculated using the whole dataset including unclosed ones as well.

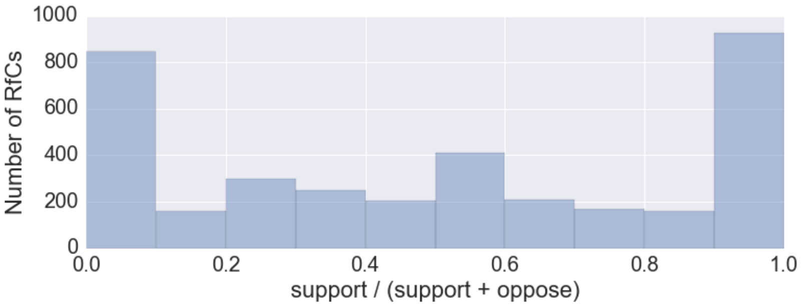


Figure 3-9: Ratio of support votes among all votes in RfCs that contain a binary poll.

it comes to using RfCs as a means to attract outside input, we find that they appear to work reasonably well. On average, 56.5% of the participants of an RfC are newcomers to the topic of the RfC, determined by considering whether the participant had previously made any edits on the talk page where the RfC took place. However, participants are relatively less experienced than initiators or closers, as shown in Table 3.6.

Discussion: A discussion's size and shape can affect both the reading and commenting experience. RfCs in our dataset had on average 34.37 comments between 11.79 participants. As a sign of how unwieldy these discussions can get, the highest number of comments on an RfC is 2,375, while the highest number of participants is 831. Both values come from the same RfC[12]. Not only can there be many comments but they can create long threads of replies. On average across RfCs, the depth of the longest thread in the discussion was 5.15 comments, while the average depth of any comment was 0.39, where a comment that is not a reply to any other comment has a depth of 0. This suggests that RfCs have a mix of deeper back-and-forth discussion as well as many comments simply responding to the initial prompt. Some of these non-threaded comments may come from a dedicated polling section within the RfC. We found that 49.6% of the RfCs in our dataset had an area for a poll. Among RfCs where there was a binary decision, on average there were 5.09 supports and 4.57 opposes, and most polls have a ratio strongly in one direction or the other (Figure 3-9).

[12] https://en.wikipedia.org/wiki/Wikipedia:VisualEditor/Default_State_RFC

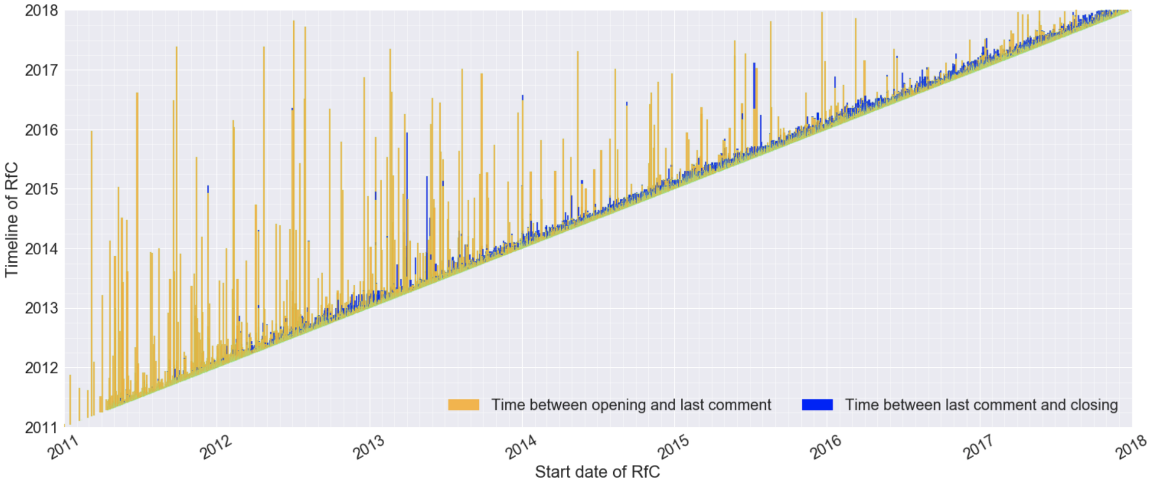


Figure 3-10: Timeline of all RfCs showing the length of time for discussion of an RfC after opening it, as well as the length of time between the last comment and the formal close, if it exists. For each RFC, we draw a vertical line whose x coordinate is the start date and whose y coordinate ranges between start and end date.

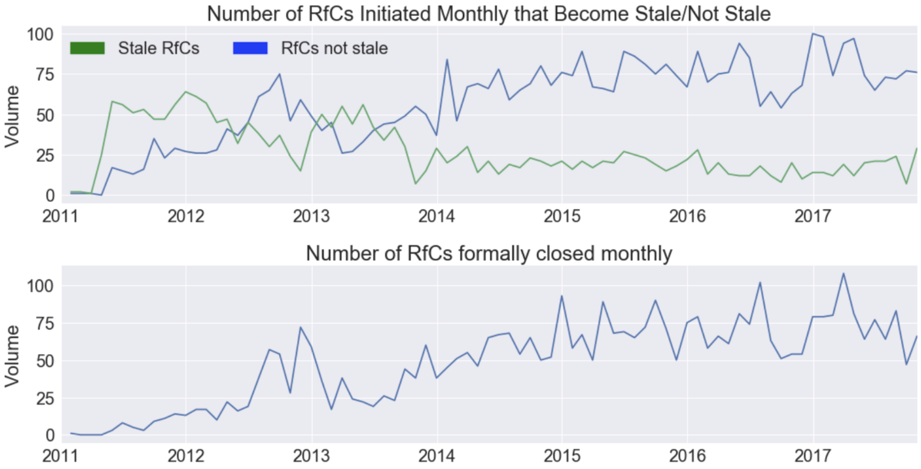


Figure 3-11: On the top, the number of RfCs initialized per month is broken down into RfCs that became stale versus RfCs that were either informally ended or formally closed. The number of RfCs formally closed each month is on the bottom.

When we calculated the length of the discussion period, we found that the average time between the first comment and the last recorded comment was 44.44 days, with a standard deviation of 160.16 days due to a heavy tail of RfCs that drag on for many months. As noted in our data collection, this duration distribution does include RfCs that were open at the time of this writing. It is also possible that at a future point in time, an editor may reopen any unclosed RfC. When considering only RfCs that were closed, the average length of the discussion was 28.17 days ($\sigma = 75.37$). In Figure 3-10, we plot the timeline of all RfCs in our dataset, with the yellow lines representing the discussion period and the blue lines representing the time from the last comment to the closing of the RfC if formally closed. As can be seen, there are many discussions that drag on for long periods of time, even years. On average, after the initial proposal, it takes 16.47 days ($\sigma = 76.89$) for the first comment to be made. This is due again to a long tail, and thus the median is 3.91 days.

Closure and Conclusion: As visualized in Figure 3-10, the time taken to close a discussion can also be long. For RfCs that eventually were formally closed, on average it took 16.74 days ($\sigma = 25.90$) after the last comment in the RfC. In total, the average RfC time period from initiation to closure for RfCs that were formally closed was 45.56 days ($\sigma = 81.14$). This is about 1.5 times longer than the default 30 days that Legobot allots, with 37% of the time spent on waiting for the closing statement.

As seen in Table 3.6, closers make up the most experienced but also smallest population, with 23% administrators. From analyzing the closer population over time, we found that the number of active closers has generally been rising since 2011. However, this population is also skewed, with 57% of the 759 closers having only closed one RfC, while the account with the most number of closes has closed 352 RfCs.

Post-Close Review: While there are no ways to automatically track what happens to an RfC after conclusion, there is a manually curated page of RfC closure reviews primarily maintained by two editors. It contains 80 RfCs from 2011 to mid-2017, representing 1.1% of the RfCs in our dataset. Of these, 40% of the closes were upheld, and 25% were changed by either being withdrawn, overturned, reverted, or reopened.

3.6.5 Why Do RfCs Not Get Closed?

Through quantitative analysis of our RfC dataset, we found a significant number of RfCs—almost half—that do not get formally closed, with about 78% of those going stale and about 22% ended informally. This can be a problem as editors involved in the RfC may be waiting on the outcome before they feel they can continue editing. It can also be discouraging if an RfC never gets closed when editors put effort into participating in the RfC. We also saw that RfCs can linger for weeks and sometimes months before getting closed, which can be problematic if the discussion has gone out of date in that time.

Data Collection

To understand why RfCs do not get closed, we conducted semi-structured interviews with 10 of the most frequent closers on English Wikipedia. In order to find interviewees, we compiled a list of frequent closers. As we did not have a dataset of RfCs yet, we instead scraped the archives of Wikipedia's Administrator's Noticeboard/Requests for closure, a board dedicated to finding closers for an RfC. This yielded links to 2,034 RfCs. We contacted 17 editors who were the most frequent closers and still active on Wikipedia, with 10 accepting.

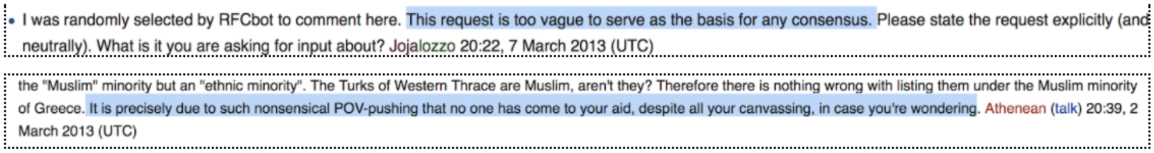
The interviews were conducted over phone or video call, with the exception of two that were conducted over back-and-forth emails. For the calls, the interviews lasted anywhere from 45 minutes to 1 hour and 30 minutes. Interviewees were compensated $15 for their time. Due to their desire for anonymity, we only have demographic information for 4 of the 10 interviewees. The average age for the four is 40.75, and all four are male. On average, interviewees have been editors on Wikipedia for 9.9 years, with only 2 of 10 with an edit history under 5 years. 3 out of 10 are administrators.

After asking general questions about interviewees' experience with RfCs, we asked interviewees to walk through the process they go through to decide what RfCs to close and how they go about closing an RfC. We asked them to consider if there were any problems with the RfC process and whether any tools or collaborations could help make the process easier or faster.

Interviews were conducted by the first and second authors. After each interview, it was transcribed and coded by them using a grounded theory approach [39] due to the exploratory nature of the study. As interviews were ongoing, the codes were discussed by all authors and grouped into major themes, including around common concerns about the RfC process as a whole and reasons for why RfCs go stale.

We also randomly selected 40 RfCs from our dataset that did not get closed and manually inspected and coded the discussion to understand why they were never closed. This analysis was coded by the first and second authors and then discussed by all the authors. Since the reasons may not always be immediately apparent from the discussion, the reasons we were able to identify were informed by our prior discussions with interviewees as well as informal conversations with top RfC participants on Wikipedia.

Problems with Initiators and Initial Proposals



- I was randomly selected by RFCbot to comment here. This request is too vague to serve as the basis for any consensus. Please state the request explicitly (and neutrally). What is it you are asking for input about? Jojalozzo 20:22, 7 March 2013 (UTC)

the "Muslim" minority but an "ethnic minority". The Turks of Western Thrace are Muslim, aren't they? Therefore there is nothing wrong with listing them under the Muslim minority of Greece. It is precisely due to such nonsensical POV-pushing that no one has come to your aid, despite all your canvassing, in case you're wondering. Athenean (talk) 20:39, 2 March 2013 (UTC)

Figure 3-12: The first meta-comment points out the initial proposal is too vague while the second notes the initiator's biased actions.

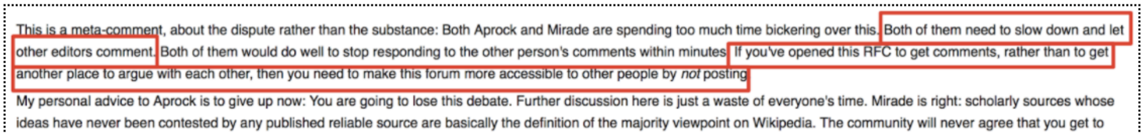
According to our random sample, issues with initiators had a lot to do with producing unclosed RfCs. 14 out of the 22 RfCs with a meta-comment had an issue related to initiator actions. For instance, sometimes the initiator was not clear with the wording of the request, potentially related to their level of experience. On the other hand, there were more severe cases when the initiator went against the normal consensus decision-making process by biasing the wording of their initial proposal or attempting to canvass by soliciting participation in a non-neutral way, either in their wording or recruitment of certain editors. A few of our interviewees (2/10) mentioned issues with initiators, with one interviewee saying:

> "*An RfC not well-formed—this can happen when the results are unclear because of the structure of the RfC. For example, the RfC might have no clear question...*"

This closer went on to say that despite this issue, it can still be possible for a closer to determine editors' opinions and make a deliberation on what editors actually ended up talking about.

Behavior of Participants: Bickering and Sock-Puppeting

Four of the RfCs that we examined explicitly mentioned excessive participant bickering, including by the initiator sometimes, which led to more complicated and longer threads that were difficult for newcomers and potential closers to examine. The back-and-forth argumentation was often caused by participants who had a history with each other and had been involved in previous discussions.



This is a meta-comment, about the dispute rather than the substance: Both Aprock and Mirade are spending too much time bickering over this. Both of them need to slow down and let other editors comment. Both of them would do well to stop responding to the other person's comments within minutes. If you've opened this RFC to get comments, rather than to get another place to argue with each other, then you need to make this forum more accessible to other people by *not* posting

My personal advice to Aprock is to give up now: You are going to lose this debate. Further discussion here is just a waste of everyone's time. Mirade is right: scholarly sources whose ideas have never been contested by any published reliable source are basically the definition of the majority viewpoint on Wikipedia. The community will never agree that you get to

Figure 3-13: Meta-comment revealing that the participants' bickering is making it difficult for other new participants to engage in the RfC.

Three of the frequent closers we interviewed also pointed out that RfCs with lots of bickering would be unlikely to become closed. One interviewee said:

> *"..no one really cares about [the RfC] that just gets a lot of bickering back and forth without a lot of substantive discussion. That's the kind of RFC that will often sit for a few months."*

Another interviewee described how excessive bickering between a few participants might also push away future potential participants:

> *"If one or two participants are trying to reply to everyone who disagrees with them, others may simply not be taking them seriously or have grown tired of repeating themselves."*

Three of the interviewees also mentioned actions by participants that try to influence the outcome of the decision by creating multiple fake accounts to create the appearance of consensus (called "sock-puppeting") or by recruiting editors to join a discussion on behalf of that editor (called "meat-puppeting" if recruiting off-wiki and "canvassing" on-wiki). When this happens and another editor notices, an investigation can be called, and the offending editor is routed to formal processes for user conduct. One frequent closer said:

> *"If I would have a suspicion that there was socking going on, I probably wouldn't be closing it."*

This was also a reason why several interviewees spoke strongly about how RfCs should not become a voting process, and mentioned that they give less attention to votes that do not include any rationale or are not based in existing policies due to these concerns.

Obvious Consensus

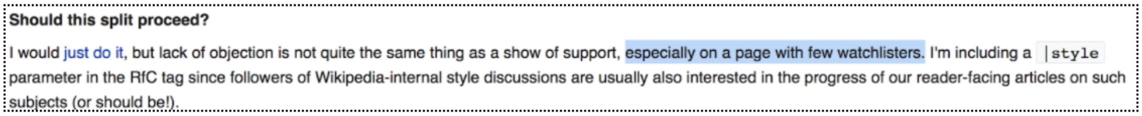
There were also cases when the outcome was an absolute consensus, and the participants seemed to think there was no need for a closure. 4 RfCs that we examined were in this category. In these cases, after numerous comments all on one side, eventually a participant just takes the RfC tag off (2/4). The other two RfCs had the tag taken off by a bot, where the participants may have just left the RfC after seeing consensus. Interviewees that mentioned this (2/10) also mentioned that many of these cases are fine to just informally end:

> *" When you have an RFC that has 15 people in support of something and one very loud person opposing it, those are very clear cut outcomes usually and it doesn't necessarily need formal closure".*

If an initiator is repeatedly starting RfCs to fight a general consensus, they may get referred to a user conduct forum. This category also included cases we saw when many participants responded to the initiator that there is no need for the RfC to begin with, which could be chalked up to lack of initiator expertise.

Lack of Interest or Expertise from Uninvolved Editors

Other than the three reasons mentioned above that were explicitly mentioned, there were also times where the reason was not clear from the discussion. Among these 18 RfCs without explicit comments about the RfC, we saw both long and short discussions. One possible reason why they did not get closed could be that there was simply lack of interest in the RfC from uninvolved editors. We noticed even in the long discussions, participants were primarily those that were already involved in the discussion before the RfC began.



Should this split proceed?

I would just do it, but lack of objection is not quite the same thing as a show of support, especially on a page with few watchlisters. I'm including a `|style` parameter in the RfC tag since followers of Wikipedia-internal style discussions are usually also interested in the progress of our reader-facing articles on such subjects (or should be!).

Figure 3-14: Comment revealing that the lack of overall interest on the page which may influence the outcome of the RfC.

Two of our interviewees also brought this up as a reason why RfCs in topics that attract only a small number of editors might go stale. One interviewee mentioned his own lack of interest in a topic being a factor, saying:

> *"When no one cares enough because even if you get it wrong, you've affected one small part of one article that might get 15 views a day, or whatever...I've definitely passed on an RFC because I thought 'this doesn't matter. My time is better used elsewhere."'*

A related issue that several closers (6/10) brought up was lack of expertise in the topic behind the RfC. While closers do not need to be experts on a topic to close it, and in fact should not be too involved in the topic so that they maintain neutrality, they still need to have some knowledge of it or be willing to invest time to learn about it. One interviewee said:

> *"...in some cases a certain amount of background may also be a requirement. This is especially relevant for more technical subjects, such as the sciences... You may be able to remedy this by studying, or it may be better to leave the discussion for someone else to close."*

And although anyone on Wikipedia can close an RfC, if the topic is too esoteric to the majority of frequent closers, then it may never get closed.

RfC is Too Complicated or Too Contentious

Two other reasons that we were not able to uncover by analyzing RfCs using meta-comments but that were mentioned by several interviewees were RfCs that were too complicated or contentious, with these problems often overlapping. Although there were no meta-comments, we noticed two long discussions containing 136 comments and 84 comments. Three interviewees mentioned that when the RfC is hard to close due to severe contentiousness, they tend to leave it to other closers who can handle it, mostly ones they felt had more authority. One interviewee said:

> "*There were a few that I avoid just because I look at it and think, 'Whoa, no way.' Usually it's the policies and guidelines, anything with like 300 plus comments or where feelings are running very high. Eventually I...think 'Hmm. That needs one of Wikipedia's big names to close."'*

Another closer mentioned that they could tell that for some RfCs, no matter how they close it, participants will follow them to their user talk page to question the close, and so they just didn't want to bother.

Other interviewees (6/10) talked about RfCs that were just too complicated to make sense of. These could be RfCs that were contentious but could also include ones that had a great deal of back-and-forth or many participants, a lot of links to outside sources or relevant policy, or a particularly content-heavy topic. One interviewee described it as:

> *"And I tried to read it, I looked it over and I realized I couldn't make heads or tails of it."*

In these cases, an RfC could stay open indefinitely if no closer wants to take on the time to make sense of the discussion and all relevant materials. We also noticed from talking to closers that most of them cited spending on the order of several hours, sometimes over the course of multiple days, closing their most complicated RfCs.

Interpersonal Issues and "Wikipolitics"

As closers are humans, interpersonal reasons also had to be considered for closures. Two RfC closers mentioned that they do not close RfCs that are related to participants with whom they have a negative relationship. Although this is not a direct reason for staleness overall, it implies that an RfC with an involved editor that has many negative relationships with other editors is more likely to stay open. One interviewee said:

> *"...my relationship with some of the contributors...is not very good. Now suppose people with whom I do not share a particularly good relationship...has initiated the RFC, I don't generally close it."*

Related to this as well as to the previous reason of an RfC being too complicated, two interviewees discussed how "wikipolitics" play into their decision to close an RfC. One interviewee said:

> "*I closed a discussion where these two people were fighting and they represented two huge factions on Wikipedia...because I did that, if you read my request for [role], that was one of the key points that people opposed it...if you have people who don't like something you did, even if you did something according to policy, if it's not popular amongst enough people, they can join their voice with something else and sway a discussion."*

For this reason, a potential closer interested in growing their social capital might steer away from the more contentious discussions.

3.6.6 Predicting the Likelihood of an RfC Going Stale

Building on our analyses of the factors related to closure, we used the RfC dataset we collected to develop classifiers to predict the likelihood of an RfC going stale. Our prediction task is framed as a binary classification problem, taking into account features related to the initiation and unfolding discussion in the RfC as well as characteristics about the article or policy page in question. We first classify RfCs into formally or informally closed versus stale using all the historical data we have on each RfC, minus the closing statement if it exists, to learn what features distinguish stale RfCs. We then consider how a model for predicting the likelihood of an RfC going stale performs as the RfC's life-cycle moves forward in time from initiation.

We used four classification algorithms and compare the performance. The four algorithms are Logistic Regression (LR), Adaptive Boosted Decision Trees (ADT), Random Forests (RF), and Support Vector Machines (SVM) with a radial-basis function kernel. We conduct training and testing on 7,087 RfCs using 61 features. For features with missing data, such as deleted user accounts, we used imputation[13] to insert the mean value instead. 50 trials were conducted with random 40% testing splits, and the resulting performance values were averaged. We also used a tree-based feature selection algorithm to find the most important features, shown in Table 3.9 based on the feature importance calculated by the ADT model. To determine feature importance we calculated Gini Importance (I) which is the normalized total reduction of the criteria due to the feature.

Features

Initiator Experience: From the interviews, we learned that initiators may have a large impact on producing RfCs that do not get closed due to lack of experience. For this reason, we calculate measures related to initiator expertise before the RfC took place, such as the *initiator edit count* and *age of the initiator account* in days. The initiator might also be well versed in Wikipedia but a newcomer to the discussion

[13] `http://scikit-learn.org/dev/modules/impute.html`

around the topic in question. Thus, we also calculate the *number of revisions to the talk page of the RfC by the initiator*. We finally considered *whether the initiator is an administrator*.

Participant Interest: Another aspect related to likelihood of closure was the ability to attract outside participation towards the RfC, which is the main goal of RfCs to begin with. Thus, we calculate the overall *number of participants* in the discussion so far, as well the *ratio of new participants* so far, where a new participant is one that has not participated on the talk page prior to the RfC.

Participant Experience: In addition to attracting participants, we saw that it was also important that participants have experience. First, an RfC that failed to attract experienced editors may be a factor in lack of interest from frequent closers, who are often also experienced editors. Experienced editors also bring a knowledge of policy and norms, potentially contributing to the quality of the discourse. Finally, sock-puppeting was noted as an issue affecting closure. This could potentially be determined by an unusually low level of experience from participants. We calculate a number of measures related to participant expertise, including the *age of the account of participants*, incorporating the average, standard deviation, sum, and maximum over those values, as well as the *participant edit count*, incorporating the average and sum.

Size and Shape of Discussion: We also found that the size and complexity of the discussion was related to the likelihood of closure. RfCs that generate a lot of discussion may have higher than usual interest and perhaps importance to the community, leading to a vested interest in closure. At the same time, these discussions might scare away potential closers who do not want to invest the time or do not feel like they have the authority. On the other hand, RfCs with very few comments may suggest lack of interest in the topic at hand. To capture these characteristics of both volume and complexity, we measure the *number of comments*, *average depth of replies* per comment, and the *average number of replies* to each comment.

Contentiousness: We learned from the interviews that a discussion's contentiousness is an important factor considered when deciding to avoid closing a discussion.

To measure this, we calculated, for RfCs that had binary polls, *number of supports/opposes*, *ratio of supports over total votes*, and average and sum of *number of replies that support/oppose comments receive*. We also calculated *weighted reciprocity*, which is a measure of the degree of back-and-forth between participants [312].

Tone of Participant Discourse: Bickering was a separate concern that was mentioned in interviews. To get a sense of the tenor of conversations, we calculated features using the frequency of terms taken from commonly used lexicons (indicative word sets) from the Linguistic Inquiry and Word Count (LIWC) software [262]. We examined the average frequency of indicative words over all comments in the discussion so far. First, we considered negative emotionality and affect, using dictionaries for *hostility*, *swear words*, and *anger*, as well as *positive* affect, *negative* affect, and *affect* terms in general. Conversely, we calculated measures for *cognition (cogmech)*, *percept*, and *insight*. Related to prior work on the importance of social aspects of deliberation [25], we also calculate measures for the use of *first-person singular words*, *inclusive* language, and *exclusive* language. Finally, we calculate measures for *certainty* and *tentativeness*.

Initial Proposal Tone and Length: Besides expertise of the initiator, we learned that the quality of the initial proposal can be important, such as if it is too short or has biased language. Thus, we measure the *number of words and characters* in the initial proposal. We also measure all the LIWC terms described in the prior feature category related to tone of participant discourse.

Popularity of RfC and Topic: Finally, we learned from interviews that the interest in the RfC and the underlying topic in question can be a factor. To measure popularity of the RfC, we calculated the the *number of words and characters in the RfC* so far, reasoning that longer and more comments indicate greater interest. To calculate interest in the general topic, we also included the *total number of revisions made on the talk page* where the RfC is located. We also look at more recent interest leading up to the RfC, including *number of revisions made 1 week, 2 weeks, 3 weeks, 1 month, and 2 months* prior to the initiation.

Algorithm	**Precision**	**Recall**	**F1**	**AUC**	**Accuracy**
LG	0.762	0.868	0.812	0.657	0.73
ADT	0.788	0.864	0.825	0.695	0.753
RF	0.75	0.909	0.822	0.645	0.736
SVM	0.71	0.955	0.815	0.58	0.709
Baseline (most frequent)	0.672	1	0.803	0.5	0.672

Table 3.7: Average performance of classifiers over 50 trials to predict the closure of RfCs from full data.

Results

First, we consider the performance of classifiers that make use of features calculated from all data from an RfC up to its closure, if there is one. We report accuracy, precision, recall, F1, and area under the curve (AUC) in Table 3.7. Adaptive Boosted Decision Trees perform the best overall except for the recall score. They achieve 75.3% accuracy while Support Vector Machines with a radial-basis function kernel perform the worst with 70.9% accuracy. The best accuracy shows a 8.1% increase over the baseline performance of 67.2% of simply picking closed for an RfC's outcome.

In Table 3.8 we report precision, recall, F1, AUC, and accuracy for an ADT classifier when using features from only one category at a time. Additionally, in Table 3.9, we show the top 14 features among all 61 features using ADT. Overall, we see that features related to *size and shape of the discussion* best model the data to predict closure, with all three features appearing in the top 14 features. Interestingly, *average number of replies* positively correlated with closure while *number of comments* and *average reply depth of comments* negatively correlated. This may be because longer depth and more comments signify greater complexity and back-and-forth arguing, which may turn some closers off. However, a greater number of replies as opposed to just one-off comments may signal greater interest in the discussion.

Another feature category that models the data well is *participant experience*, with features related to the Wikipedia age of and number of edits by participants listed as important. All of these features were positively correlated with closure, indicating the importance of experienced participants.

Category	**Precision**	**Recall**	**F1**	**AUC**	**Accuracy**
Size and Shape of Discussion	0.750	0.903	0.819	0.644	0.733
Participant Experience	0.757	0.860	0.805	0.647	0.720
Participant Interest	0.722	0.897	0.800	0.595	0.699
Contentiousness	0.674	0.980	0.799	0.506	0.669
Popularity of RfC and Topic	0.687	0.947	0.797	0.533	0.675
Tone of Discourse	0.691	0.925	0.791	0.54	0.673
Initiator Experience	0.675	0.984	0.801	0.508	0.672
Initial Proposal Tone and Length	0.673	0.978	0.798	0.504	0.667

Table 3.8: Performance of ADT classifier to predict the closure of RfCs using features from each category.

Features	**Importance**	ρ	$\boldsymbol{p}$
Number of comments	0.08	-0.053	< 0.0001
Maximum Wikipedia age of participants	0.06	0.12	< 0.0001
Cognitive tone of RfC	0.06	-0.049	< 0.0001
Average Wikipedia age of participants	0.06	0.003	< 1
σ of Wikipedia age of participants	0.04	0.215	< 0.0001
Sum of edit counts of participants	0.04	0.147	< 0.0001
Average edit counts of participants	0.04	0.146	< 0.0001
Number of participants	0.04	0.13	< 0.0001
Average reply depth of comments	0.04	-0.13	< 0.0001
Average number of replies	0.04	0.061	< 0.0001
Affective tone of RfC	0.04	-0.054	< 0.0001
Wikipedia age of RfC initiator	0.04	0.028	< 0.05
Hostile tone of initial proposal	0.04	0.013	< 0.5
First person singular word usage of RfC	0.04	0.015	<0.5

Table 3.9: Top 14 features in the ADT model incorporating all data, including correlation to closure.

While not performing as well altogether, a few features related to *tone of participant discourse* and *tone of initial proposal* were included in the top 14 features. For instance, the affective tone of the discussion was weakly negatively correlated with closure, possibly because words related to emotion may hinder progress of a deliberative discussion.

Lastly, *Wikipedia age of RfC initiator* was also included in the top 14 features with a weak positive correlation with closure. This implies a higher level of an initiator's expertise may help prevent an RfC from going stale.

Predicting closure as RfCs progress While we demonstrated that we can classify closed versus unclosed RfCs from our dataset when provided with all the RfC participation, a more interesting question is how soon after an RfC is initiated can we begin to predict the likelihood of closure with reasonable accuracy. To understand this, we built models that predict closure at different points in time after the start of an RfC. Immediately after initiation, features from the categories of *initiator experience*, *initial proposal's tone and length*, and *popularity of RfC and topic* can be used. As time goes by and participants join the conversation, we can make new predictions about the likelihood of closure using all 61 features and updating their values with historical data.

As time moves forward from initiation, we perform a prediction each week. However, some RfCs get closed during that time—since we already know the outcome of those RfCs looking back in time, we can discard already-closed RfCs in each week's prediction. This means that at each week, we only make predictions on the RfCs that are as of yet unclosed. Since as time goes by, some unclosed RfCs may start to go stale as there are no new comments, we also add a time-based feature to these models which is *the number of days since the last comment up to the current point in time*. We choose to do this instead of discarding inactive RfCs from our prediction since any unclosed RfC might be re-opened at any time by an editor, and this is unknown ahead of time.

As the accuracy over time in Figure 3-15 shows, all four classifiers start out quite close to a baseline which simply predicts closure for all RfCs, achieving around 66%

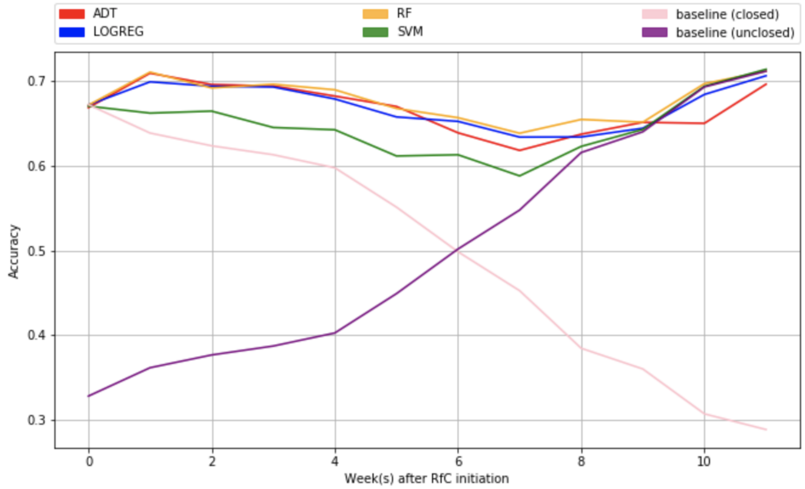


Figure 3-15: Change in accuracy over time after initiation up to 11 weeks.

accuracy. However, as time moves forward across RfCs and only unclosed RfCs remain, the baseline for simply predicting closure for all remaining RfCs drops while the baseline for simply predicting going stale improves. Similarly, as time progresses, the accuracy of the classifiers begin to approach the value presented above with all the RfC participation data baked in, demonstrating how our models can provide timely feedback to participants even just a week after the RfC is initiated. As time goes to 11 weeks after initiation, the baseline prediction of marking all RfCs unclosed begins to approach our models' performances, as most RfCs that are still unclosed at this point are likely to go stale.

3.6.7 Design Implications

Through a comprehensive analysis of RfCs on English Wikipedia, we examined how RfCs get initiated, discussed, and closed. We found that while the closer population and the proportion of RfCs getting closed is increasing over the last seven years, a large portion of RfCs still do not get closed in a timely manner. From interviews and qualitative analysis of unclosed RfCs, we notice various factors including the nature of discourse and the characteristics and number of discussion participants can indicate the likelihood of resolution. Using measures informed by interviews and inspection of RfCs, we were able to develop a model that can predict the likelihood of closure at above 70% even a single week after initiation of the RfC. These suggest design considerations for tools that could potentially help make formal deliberations on Wikipedia more effective.

Tools to help initiators and participants

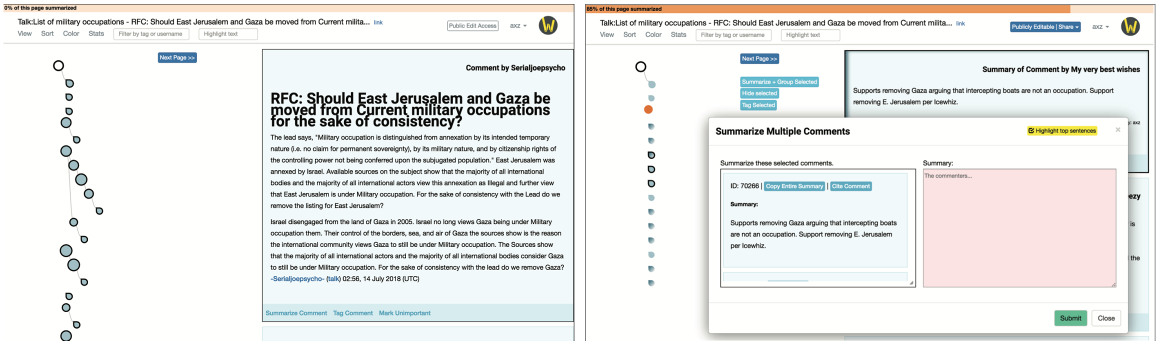
First, our development of a model for predicting closure could be helpful as a tool for initiators or participants in an RfC to consider ways to avoid going stale. From the model utilizing all participation data, we find that the *participants' interest and experience* were some of the most important factors. In terms of participants' *interest*, it seems crucial to find a way to properly promote an RfC to experienced Wikipedians. Although we did not include it in this work, it would be interesting to find what are the most effective ways to gather interest in an RfC. For example, it might be effective for certain topics to publicize an RfC in particular forums within Wikipedia. Or perhaps certain ways of phrasing the solicitation for participation or closure makes a difference. This kind of feedback, in addition to the feedback that our existing model provides, could help suggest actions for users to take when waiting for more participants or a closer.

As the results imply that participants' expertise is crucial for an RfC to become resolved, this demonstrates the need for designs that can provide editors with relatively lower level of expertise to communicate or receive feedback from more experienced participants. As an interviewee mentioned, participants learn how to provide more reliable sources and policies as evidence by observing or even being won over by more experienced editors' comments during deliberation. A system that can match and invite a group of experienced editors to an RfC that has relatively inexperienced participants could be helpful. Future work could analyze the Feedback Request Service, one of the primary drivers for soliciting participants, to consider whether alternative designs such as pings to volunteers that are not simply random or that happen at different points in the RfC's life-cycle could be beneficial. This is also the case for helping out initiators when writing the proposal, as the initiator's experience was the most crucial factor at the time when one is initiating an RfC.

Tools to help closers

In addition, we learned about how the *size and shape of discussions* is predictive of going stale. This finding echoes interviewee responses that mentioned spending hours combing through long and deep discussions before writing a resolution, as well as sometimes purposefully shying away from RfCs that were too complicated or contentious. This suggests that tools like Wikum [385] to better parse and organize these long threaded discussions could potentially help manage the workload. A complementary direction could be to consider how similar tools could facilitate closing larger RfCs collaboratively as opposed to by a single individual. While frequent closers tell us that these do happen on rare occasion in Wikipedia on an ad hoc basis, they generally involve collaborations over the draft closing statement through back-and-forth email as opposed to collaboratively understanding and organizing a massive discussion. Additionally, by sharing responsibility it might lesson concerns about "wikipolitics" or lack of authority.

It would also be interesting to consider ways that participants in a deliberation could enrich the representation of the discussion to provide more information that can help closers. For instance, sites like Reddit's ChangeMyView allow discussants to mark when a particular argument has changed their mind on a topic. Since RfCs are meant to be consensus-driven as opposed to voting-based, the deliberation should ideally be causing people to come together over time. Illuminating points of consensus and persuasive arguments would be helpful to closers and may speed up consensus since new participants will more quickly get up to speed. Similarly, an idea that a frequent closer mentioned was a tool to allow one to see the RfC discussion unfolding over time, so that he could notice changes in people's interest and opinions as time went on. Currently, he achieved this by going through the revisions on the RfC page by page, which he found to be tedious.


0% of this page summarized
Talk:List of military occupations - RFC: Should East Jerusalem and Gaza be moved from Current milita... link
View
Sort
Color
Stats
Filter by tag or username
Highlight text
Public Edit Access
axz
Next Page >>
Summarize Comment
Tag Comment
Mark Unimportant
85% of this page summarized
Publicly Editable | Share
Summarize + Group Selected
Hide selected
Tag Selected
Highlight top sentences
Copy Entire Summary
Cite Comment
The commenters...
Submit
Close

Comment by Serialjoepsycho

RFC: Should East Jerusalem and Gaza be moved from Current military occupations for the sake of consistency?

The lead says, "Military occupation is distinguished from annexation by its intended temporary nature (i.e. no claim for permanent sovereignty), by its military nature, and by citizenship rights of the controlling power not being conferred upon the subjugated population." East Jerusalem was annexed by Israel. Available sources on the subject show that the majority of all international bodies and the majority of all international actors view this annexation as Illegal and further view that East Jerusalem is under Military occupation. For the sake of consistency with the Lead do we remove the listing for East Jerusalem?

Israel disengaged from the land of Gaza in 2005. Israel no long views Gaza being under Military occupation them. Their control of the borders, sea, and air of Gaza the sources show is the reason the international community views Gaza to still be under Military occupation. The Sources show that the majority of all international actors and the majority of all international bodies consider Gaza to still be under Military occupation. For the sake of consistency with the lead do we remove Gaza? -Serialjoepsycho- (talk) 02:56, 14 July 2018 (UTC)

Summary of Comment by My very best wishes

Supports removing Gaza arguing that intercepting boats are not an occupation. Support removing E. Jerusalem per Icewhiz.

Summarize Multiple Comments

Summarize these selected comments.

ID: 70266 |

Summary:

Supports removing Gaza arguing that intercepting boats are not an occupation. Support removing E. Jerusalem per Icewhiz.

Summary:

Figure 3-16: On the left, the original deliberation taken from Wikipedia and imported into Wikum. On the right, the same discussion, now partially summarized by a field study participant. An editing modal is opened, demonstrating the tool for summarizing a group of comments.

3.7 Field Study of Wikum on Wikipedia

From studying RfCs and frequent Wikipedia closers [154], we learned about their existing workflow for closing deliberations, finding that closers often spend hours in one sitting on a single closing and use few aids other than basic note-taking tools to keep track of their work. This suggests that Wikum could potentially be useful, both individually and collectively, to help closers.

We conducted a field study of the tool with 8 frequent Wikipedia closers, where they voluntarily used the tool to formally close an open discussion on Wikipedia. We found that the tool was particularly effective for open-ended discourse, and that the task breakdown facilitated by the tool reduced cognitive load for participants, allowing them to split work between multiple sittings. It was also easier for participants to switch back and forth between lower-complexity tasks like tagging and higher-complexity tasks like summarization. Finally, we found evidence from two participants that the tool can help break up work towards collaborative closures through the emergence of user roles.

3.7.1 Existing Workflow for Frequent RfC Closers

From the interviews conducted with frequent closers about why RfCs do not get closed in the prior section, we also asked closers to talk about their process for closing discussions. The method for conducting and analyzing interview data was the same.

Reading the Discussion: Although the closing process varied slightly from person to person, most of the interviewees answered that the process started with fully reading the discussion (9/10) while one interviewee preferred reading the RfC question and related topics including sources (1/10). Interviewees mentioned that often times the consensus is clear after the first read-through. When the discussion is more difficult and complicated, interviewees would read the discussion several times and more carefully weigh the reasons of each side before writing a closing statement.

Tools: Many of the closers (8/10) replied they used tools like Notepad to jot down notes about participants' arguments when the deliberation is complicated. One interviewee described manipulating the discussion in Word:

> "*If an RfC is especially complicated, I'll copy the text into Word and simplify the discussion...I'll break up walls of text, delete comments..., group comments in a way that's more logical to me, order comments by strength of argument, etc.*"

Another interviewee described using Wikipedia's edit history to page through individual edits so as to not see the entire discussion at once. However, most interviewees did not use tools heavily, and two stated they kept everything in their head.

Closing Statement: Time spent on writing the closing statement seemed to vary, with one interviewee mentioning:

> "*Once I'm done with the examinations, writing a result is usually quick and easy.*"

However, another interviewee said:

> "*The writing part is always about one to one and a half hours, because there are parts of drafting, there are parts of, again, reading some parts*

> *which I may have necessarily forgotten...So writing in general takes more time than reading*".

The preferred style of writing the closing statement seemed to vary as well, with some interviewees preferring concise statements while others thought more detailed closings were clear and helpful to other editors.

Time Spent: Five interviewees mentioned that they prefer to close discussions in one sitting, with one person saying:

> "*I find when I split them up, I forget what I've done last time*".

However, most interviewees recalled times when they had to close large discussions that took many hours, sometimes requiring them to break up the work into multiple sessions:

> "*Sometimes it takes four or five hours and you don't have that time.*"

Considerations When Closing

Interviewees mentioned that they look for *consensus* when closing RfCs. Based on the interviews, we discovered that consensus can be decided by multiple factors, meaning closers work to take various considerations in mind. Here we describe the most important considerations when closing.

Consensus Informed by Policy over Voting: Discussions on Wikipedia including RfCs can take various forms, with some of them containing a section for polling while others are more open-ended. In the polling sections, editors write their stance much like a vote, such as "support" or "oppose" along with the reasons. Several interviewees (6/10) felt that RfCs are not a voting process and that the closure should consider multiple factors including strength of the arguments and their adherence to existing policies. However, some other interviewees (2/10) mentioned that the majority vote was still important, nevertheless:

> "*People say it's not a voting process. It kind of is in a sense...It shouldn't be a large part of the decision but it is there and it is relevant.*"

The majority of interviewees also answered that they carefully considered Wikipedia policies during the closing process (7/10). The term *policy* seemed to include both policy and guideline pages on Wikipedia as well as unwritten norms that can only be picked up through experience. Interviewees also mentioned that considering policies is crucial when counting votes as well, since not all votes may follow policy.

Maintaining Neutrality and Clarity in Writing: Many interviewees described the importance of staying neutral while closing an RfC (7/10), which could sometimes be difficult. One interviewee said:

> "*Probably the main reason for a bad close is when someone has strong views on a subject. They may ignore arguments they dislike, and just write their own view as the result.*"

As a result, interviewees tried to avoid any potential accusations of bias from people who might disagree with their close, such as by not closing discussions where they had a strong opinion.

Some interviewees (3/10) also described how they strive to be clear and unambiguous in their closing statement. Part of the reason was because they knew that if they were not, participants would go to their user talk page to ask questions about or challenge the close.

Individual vs. Collaborative Closing

Though most of the time closing an RfC is a solitary task, in a few cases, closures are conducted collaboratively. Most of our interviewees preferred closing alone, with one interviewee saying:

> "*The thing is, though, that takes a lot of work and coordination. Multi-party closes aren't common, and they usually aren't needed. Only if the consensus is difficult to determine, and if it's an important matter that's going to affect a lot of articles.*"

4 out of the 10 interviewees said they had positive experiences with collaborative closures. They described the process mainly as one person conducting the close in

full, followed by a second or third person who also reads over the discussion, looks over the draft closing statement, and then gives feedback. Overall, they described little actual division of work:

> "*In my example, we analyzed the RfC independently then discussed it with each other. We all came to the same conclusion, except for a couple of minor points which we resolved through discussion, then we proposed drafts until we came to agreement on the wording.*"

One reason that interviewees liked them despite the duplication of work was because it helped correct for bias or missing information. An interviewee mentioned:

> "*Yeah I like joint closure, because sometimes I have found out that my views were not correct. Probably the summary that I alone made out...was not optimal, and another closer helped me reach the optimal conclusion.*"

Another interviewee mentioned that collaborative closures seemed to work well for discussions that involve large policy changes or are controversial.

3.7.2 Field Study

In the interviews, participants mentioned spending on the order of several hours towards closing especially long RfCs. We now consider how a tool for breaking down the work could help Wikipedia closers summarize and resolve large deliberative discussions more easily. While Wikum and the recursive summarization approach has been studied in lab settings in past work [385], it has not been studied in real environments. While one could imagine large groups of people working to summarize a discussion collectively using this tool, in this work, we focus primarily on individuals using Wikum to summarize a large discussion.

Study Recruitment

While Wikum is a free and open tool, we needed to publicize it to Wikipedians who have experience closing deliberations. Thus, we reached out to the same list of

	Experience on Wikipedia			Prior Activity Closing RfCs		
	Account Age (years)	Total No. Edits	Admin	No. RfCs Closed	Avg (σ) No. Comments in Closed RfCs	Avg (σ) Len Closing Statement (words)
P1	14	30,000	Y	27	50.4 (7.58)	110.15 (68.68)
P2	5	20,000	N	44	44.27 (5.99)	45.66 (29.06)
P3	9	100,000	N	43	41.67 (1.12)	25.77 (11.83)
P4	6	50,000	Y	8	23.87 (4.2)	100.5 (58.05)
P5	3	100,000	Y	58	39.41 (1.5)	76.86 (70.25)
P6	11	250,000	Y	92	34.27 (2.32)	85.11 (85.97)
P7	7	10,000	N	28	59.68 (8.44)	105.11 (109.27)
P8	14	60,000	Y	-	-	-

Table 3.10: Field study participants and information regarding their overall experience on Wikipedia. We also calculate and report their prior experience with closing RfCs. (P8 has no results because they are a frequent AfD closer but not RfC closer).

top RfC closers we created for finding interviewees. Since we noticed that many of the top closers were active at one point in time but had not closed anything in a while, we also looked for people who had recently closed an RfC, according to the Administrator's Noticeboard/Requests for Closure. While some of these people may be less experienced, their experience with closing is fresher, and they would likely be more interested in trying out the tool than people who hadn't closed in years. We alerted people from the above groups to Wikum by emailing them or posting to their User talk pages on Wikipedia. No users were compensated for participation.

In the end, we had 8 people who went on to use Wikum while closing an RfC. Their prior experience in Wikipedia as a whole as well as in closing RfCs can be seen in Table 3.10. Information about users' prior experience with closing RfCs was taken from a comprehensive dataset of RfCs recently released [154]. Most participants had extensive experience both on Wikipedia and with closing RfCs or closing deliberations in general. Indeed, two participants were current or former members of the Arbitration Committee (ArbCom), or Wikipedia's highest court or court of last resort for disputes. One user, P8, had no experience with closing RfCs but was a frequent Articles for Deletion (AfD) closer, a similar vehicle for deliberation but one that we have no statistics on. Three of the participants were also interviewed for the first part of this paper.

Study Task

We invited closers to find an open RfC that they would like to close and work on closing it with Wikum. We provided to users a written tutorial explaining the features of Wikum in a shared Google Doc file. Users were asked to create an account on Wikum and stay logged in but were otherwise free to use Wikum how they liked for as long as they liked to close the RfC. While 6 users ended up closing RfCs on their own, 2 users chose to work together to close an especially large RfC collaboratively.

Data Collection

We collected logging data from the Wikum website, including all edits to the Wikum page, which were captured and available as edit history to users, as well as additional logging of page visits, API calls, and user activity within the page, such as when they opened a modal to edit a summary.

We also reached out to users to fill out a survey and participate in an interview once we saw that they had used Wikum. In total, five of the participants completed the survey and four participated in the interview, which lasted around 30 minutes or was conducted over email. We asked participants about their experience both making sense of the discussion and summarizing using Wikum, in comparison to what they did before. We also asked participants their thoughts on the usefulness of particular features and whether they would use Wikum again and in what circumstances.

Results

Switching Between Different Tasks Over Time: In Figure 3-17, we show the progress over time by P1-P6, who used Wikum on their own to close a discussion. On the X-axis, we show time spent on Wikum, adding dashed lines to indicate when a user has gone away for a period of time. On the Y-axis, we show the percentage of the discussion that has been summarized, where 0% represents the original discussion

	No. Comm-ents in RfC	Total Time Spent	% Time Spent Summ-arizing	Times Viewed Author Info	No. Tags Made	No. Summ-aries Made	Avg Len Summ-aries (words)	Len Closing Statement (words)
P1	66	1:27:01	11.74%	4	7	18	17.39	292
P2	139	1:23:38	11.70%	5	6	13	21.83	237
P3	78	2:38:04	28.01%	34	8	28	17.58	80
P4	74	1:03:16	21.21%	13	0	41	4.35	248
P5	45	0:39:54	23.10%	2	0	7	37.00	-
P6	47	0:27:43	12.27%	4	3	13	7.25	-
P7	202	1:26:23	15.59%	19	2	34	7.21	-
P8	202	0:42:51	00.58%	0	2	5	13.2	370

Table 3.11: Field study participants and their activity during the field study. P7 and P8 participated in a collaborative closure together.

and 100% completion is when enough original comments have been summarized so that there are only 250 words to read at the outset. As can be seen, some users fall short of 100%, and indeed most actually chose to write their final closing statement on Wikipedia and not in Wikum. Through examination of how participants used features in Wikum over time, as well as post-study interviews, we gained a sense of their workflow for completing the close.

We see examples of a bottom-up sensemaking process [267] in some of the workflows, where lower-complexity tasks such as tagging, hiding, or moving is enacted on individual or small subthreads of comments, followed by the higher-complexity summarization of comments and threads at increasingly higher levels. P1 demonstrates this workflow, with a round of tagging leading to summarizing more and more of the discussion. However, we noticed participants did not always strictly move from lower-complexity tasks to higher-complexity tasks. For instance, both P3 and P6 interleave summarization with tagging and hiding comments, even towards the end of the process. And though P4 and P5 conduct mostly summarization tasks, these are spread out across the timeline, indicating a mix of reading and summarizing. By breaking a large complex task such as summarization of a large discussion into smaller complex tasks, Wikum provides the ability to ease in to each summarization action

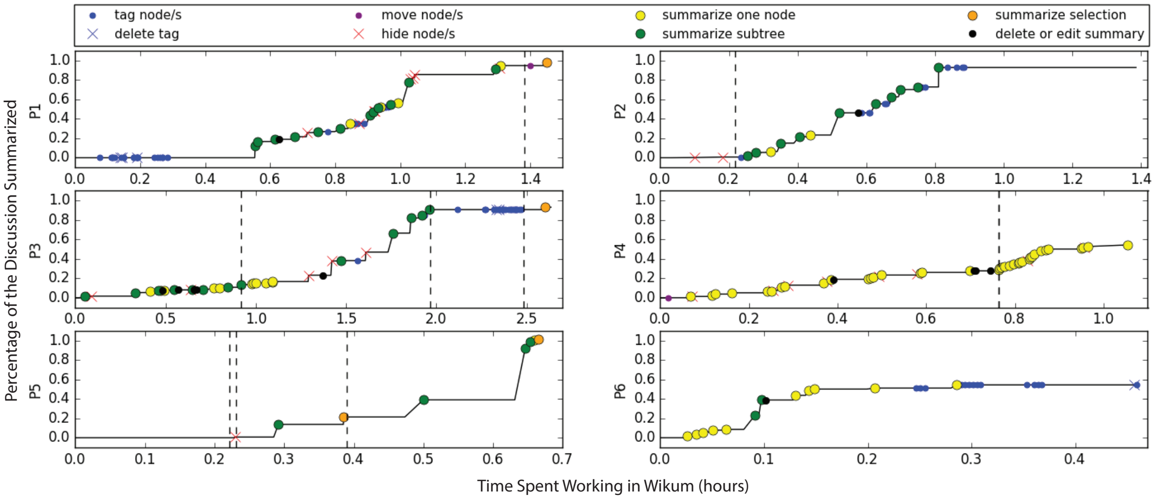


Figure 3-17: Progress over time by P1-P6. On the X-axis we have the time spent within Wikum, with dotted lines representing breaks in time. The Y-axis shows how much the original discussion has been reduced down.

with other lower-complexity work [34].

In Table 3.11, we can see that P3, P4, and P5 spent a fifth to a quarter of their time summarizing. In comparison, for P1, P2, and P6, the time spent summarizing was closer to one tenth, indicating they spent more time in tasks for organizing or cleaning, which we can see reflected in Figure 3-17. For instance, P1 described their workflow as centered around grouping similar authors and comments through the use of tagging:

> "*...I...try to categorize comments into major positions (so the 'tag' feature is great)...I wrote down the main perspectives...and then categorized people into positions. Then I would draw up my closing statement, taking inspiration from the relative sizes of each category.*"

For participants that spent more time summarizing, they mentioned that Wikum allowed them to easily interleave different types of tasks whereas before, reading and synthesizing were done separately:

> "*[Before,] I read the question, read through the entire set of comments, and try to arrive at a conclusion...Working in Wikum, I would summarize and/or mark as unimportant each comment as I read it*" (P4).

P5 made similar comments, mentioning that this ability made the overall process faster:

> "*Usually...I read through several times the entire thing and then I go back and drill down...[In Wikum] I was summarizing as I was going, it definitely made things faster...*"

Task Breakdown Reduces Cognitive Load: We found evidence during the field study that the task breakdown and ability to switch between types of tasks in Wikum helped to reduce participants' cognitive load. P5 said:

> "*It was really good...at breaking out different discussions into different pieces. I think the real...strength was that instead of having this big conglomerate of text to work through, I had these little subthreads. I could compartmentalize it a little bit better...*"

P4 made similar remarks about the summarization aspects of Wikum, while P3 mentioned how tagging within the tool helped reduce mental load as well. From the survey, users rated Wikum a 4.0 on average on a 5-point Likert scale, for reducing mental or cognitive load.

One aspect of this reduction in load meant that users could close a discussion in Wikum over multiple sittings. We noticed that several users split up their work over the course of multiple hours and days, including as much as 4 days in the case of P2 and P4. Indeed, as of this writing, P5 and P6 have yet to write their final closing statement as they felt there was no rush, and they had their summarized discussion in Wikum to work off of. This is markedly different from what our interviewees stated, where they said that they primarily conducted their close in one sitting. P3 said:

> "*I split the analysis into three sessions over three days. There was no problem coming back. Wikum helped the process because the summaries make it easy to know where you left off.*"

P5 also mentioned how ability to break down the work over multiple sittings made it easier to tackle longer discussions:

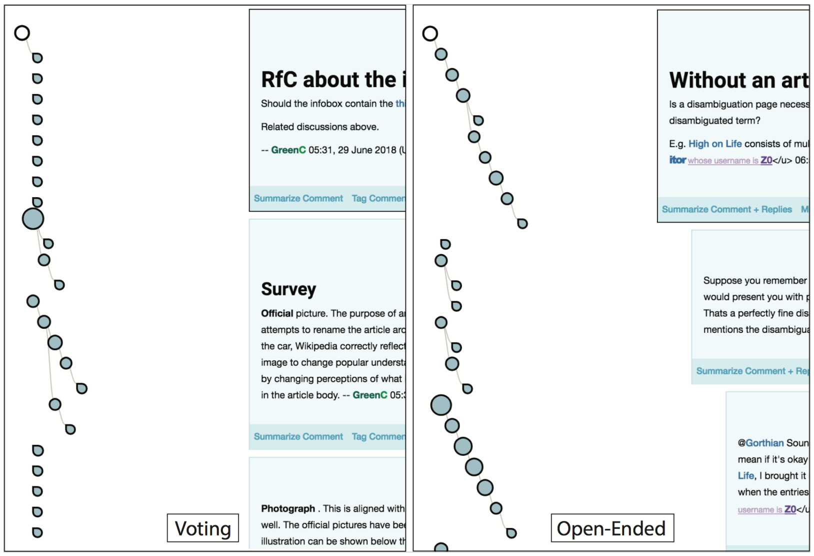

RfC about the i
Should the infobox contain the thi
Related discussions above.
-- GreenC 05:31, 29 June 2018 (U
Summarize Comment
Tag Commen
Survey
Official picture. The purpose of a
attempts to rename the article arc
the car, Wikipedia correctly reflect
image to change popular understa
by changing perceptions of what
in the article body. -- GreenC 05:
Summarize Comment
Tag Commen
Photograph . This is aligned with
well. The official pictures have be
illustration can be shown below th
Voting
Without an art
Is a disambiguation page necess
disambiguated term?
E.g. High on Life consists of mul
itor whose username is Z0</u> 06:
Summarize Comment + Replies
M
Suppose you remember
would present you with
Thats a perfectly fine dis
mentions the disambigua
Summarize Comment + Re
@Gorthian Soun
mean if it's okay
Life, I brought it
when the entries
username is Z0</u
Open-Ended

Figure 3-18: On the left is an example of a discussion with more voting from our field study in Wikum, while on the right is a more open-ended discussion.

> "*I don't usually find myself able to do that when I'm closing a big RfC, because if I were to go away and come back, I would have forgotten what I had read essentially. Whereas with Wikum, since I had the summaries, I could just read the summaries and my memory was refreshed.*"

Not only is it easier to complete a large discussion, these factors may also make the task less daunting to begin [334]. P1, when asked if they would use Wikum again, said:

> "*Probably, especially long, protracted, open discussions like the one I closed this time! I doubt I would've tackled that one on my own without a tool like Wikum.*"

Indeed, all of the RfCs self-selected by the participants had a comment count greater than one standard deviation than their average comment count over all the prior RfCs they closed, as shown in Table 3.10.

Open-ended Discussions Versus Voting Discussions: Given the reduction in cognitive load that Wikum provides, it is not surprising that users felt that Wikum was best suited for longer discussions with deeper threads. Most users also drew a distinction between discussions with a closed set of outcomes versus more open-ended discussions. An example of each is shown in Figure 3-18. P1 felt that Wikum was better for:

> "*...discussions that explore a problem and require a solution, but don't come with a pre-determined set of possible outcomes, like 'how should we reform RfA' or "how should we present [controversial topic].*"

This suggests that processes such as the Village Pump on Wikipedia, where many proposals are developed, would be a better use case for Wikum compared to processes such as Articles for Deletion (AfD), which has a great deal more voting on whether to delete an article or not. P5 also mentioned that Wikum was helpful for discussions where the question itself was not clear at the outset:

> "*Especially with this type of discussion where...the way it was handled I was very confused, there wasn't a very clear question at first...because it was so muddled, I probably did it in half the time I would have without Wikum.*"

In the surveys, we asked participants to rate the effectiveness of Wikum for making sense of the discussion as well as for summarizing the discussion, with results favoring its summarization capabilities (3.8 vs. 4.6 on average on a 5-point Likert scale). While participants rated the tagging and hiding tools highly on average (4.0 and 4.6 respectively), they were less enthusiastic about Wikum's support for automatic clustering or its tag suggestions, which use algorithms developed for generic text. To better support discussions that have more of a voting flavor, Wikum could make use of the existing wikimarkup to automatically extract, group, and tally votes. But to reduce emphasis on simply counting a majority, which goes against Wikipedia's spirit of finding consensus, the tool could also allow closers to attach different weights to different votes.

Improvements to Closing Statements for Newcomers: We received some feedback from participants that the Wikum process changed the quality or content of their final output, including the closing statement. P1 stated:

> “*My closing statement felt more like an abridged story than a judge’s ruling.*”

Indeed, we noticed in Table 3.10 that of the discussions that had a closing statement, they tended to be much longer, anywhere from 2.5 times to 5 times longer, than the typical closing statement that our participants had written in the past. When asked whether this was a positive change, P1 stated:

> “*For larger, more polarised discussions, no amount of story-telling would quell future controversy, but a story-style closing rationale might still help future readers understand the outcome of that discussion.*”

That is, while a longer closing might only sometimes be useful to the discussants, echoing remarks from the interviews, it would likely always be useful to newcomers. Both P3 and P5 mentioned that the Wikum process ensured that their work was thorough and that they could cut through the many tangents, with P5 stating:

> “*...by being able to read things in the different compartments or summarize different threads...it makes it much easier to find a consensus among all the...jibber jabber back and forth between people.*”

Collaborative Closures: Given Wikum’s ability to break down the summarization task, we considered whether it could also be used to support collaborative closures. P7 and P8 were the only participants we approached that were interested in trying out a collaborative close. Most participants we approached had little experience with closing collaborative discussions or did not want to try them for this task. Echoing findings from our interviews, P5 stated:

> “*...I think we’re very poor at dividing labor when we do collaborative closes...we basically just end up having three people closing the RfC and whatever two of them agree on is what the close is.*”

However, there are occasionally deliberations on Wikipedia that reach into the hundreds of comments and involve difficult and consequential decisions. In these

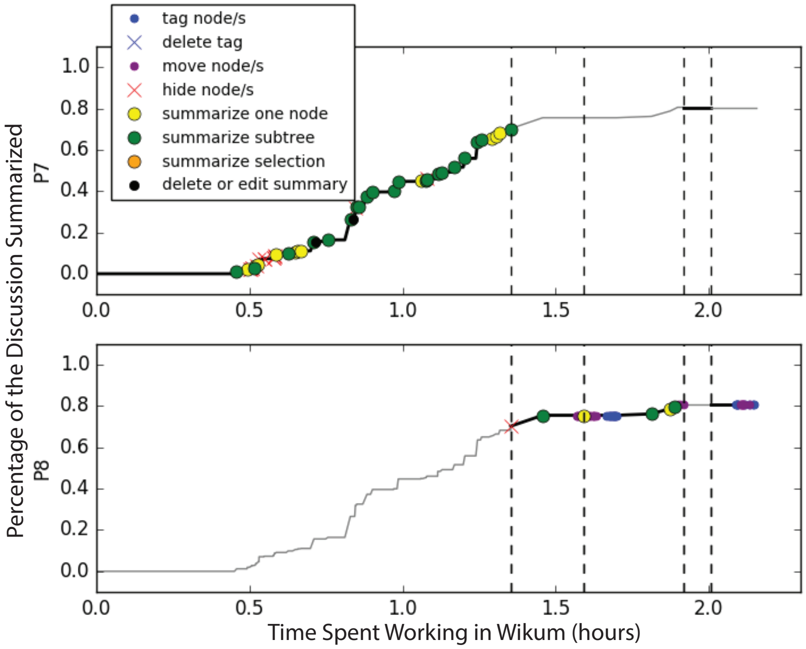


Figure 3-19: Progress over time by P7 and P8, who worked together on the same discussion.

cases, it would be particularly beneficial to break up the work if possible. For our study, P7 and P8 agreed to work on a 202 comment discussion over the course of a week and a half. We can observe their work on Wikum in Figure 3-19. Only P8 answered questions after the study.

We noticed how the two users initially settled into different roles, with P7 focusing on summarizing and P8 focusing more on organization, such as moving comments into "for" and "against" groups and tagging them. As shown in Table 3.10, P8 spent less than 1% of the time summarizing in the 43 minutes they were working on the discussion. However, in the end, P8 wrote the final closing statement (not in Wikum), as they learned from P7 that P7 had formed an opinion on the RfC over the course of summarizing.

Interviewees had previously mentioned that they would not trust the summaries written by someone else, making collaborative summarization useless. However, P8 was willing to trust someone they felt was experienced. In addition, Wikum allowed them to easily check their partner's work:

> *“I don’t think [P7] and I interacted previously but they are an experienced editor in good standing, so I assumed that the summaries were correct but I still checked them a bit. If they had been done by someone else without this status, I would have checked them all in detail, because then I had no reason to assume them to be correct.”*

While we provided no direction regarding roles, emergent behavior suggests that in the future, established roles on Wikum could allow for specialization. For instance, more junior closers could organize the discussion and close smaller chunks of discussion and more experienced closers could perform the rest of the summarization. P8 felt that breaking up tagging and summarizing would be productive:

> *“...if one person can weed out the irrelevant stuff, the rest can focus on the important parts...While tagging and summarizing can be done together, it makes sense for one person to summarize and the other to tag, so both can check whether the other agrees with the assessment, thus strengthening the result.*”

Such an arrangement might be useful given the relatively small population of closers [154] and no current established process to gain experience as a new closer. When asked whether less-experienced closers could “shadow” an experienced editor with lower-level tasks, P8 stated:

> *“I think it will likely help them to gain some experience. After all, they will do the work the closer does when assessing a discussion and learn what is and isn’t important.*”

More studies are necessary to test these ideas, though the first step is convincing more closers that it is feasible.

3.8 Discussion

3.8.1 A Tension Between Summary Goals

From the user studies of both creating and reading summaries, we learned what users perceived was useful about Wikum as well as what they desired in a summary. Some users were interested in getting an overview of the topic of the conversation, with points organized in pros and cons and grouped by topic. Other users saw summaries embedded in the discussion structure as useful signposts for readers to decide whether to go down that particular path to find interesting comments.

These two modes suggest slightly different design decisions for both readers and editors regarding whether Wikum should allow editors to break the original discussion structure in the process of grouping related comments and curating. While there are benefits to breaking discussion structure, there are possible pitfalls as well. For users more interested in following a thread of conversation, it would be important to still be able to see comments in their original context. We noticed in a pilot study that editors were reluctant to break the original discussion structure out of concern about altering original commentators' intents. However, in the Wikipedia field study, some closers often moved comments around into different categories, breaking structure. This may be because they were consider Wikum like a private workspace that no one else would see.

One open question is what workflow is best for summarizing deliberative discourse, where "best" could refer to most efficient, intuitive, or considerate of all contributions. Some users chose to focus on grouping comment authors by their stance, while others grouped comments by their vote using tagging. Still others went subthread by subthread, summarizing in a bottom-up fashion. As conveyed by our Wikipedia field study participants, the answer may depend on how constrained the discussion is at the outset (poll-like or threaded). The current recursive summarization workflow in Wikum makes it easy to follow thread organization and summarize in a bottom-up fashion and is thus more suited to open-ended discourse. However, more tools are necessary to support more top-down workflows, where categories and parameters may be defined at the outset.

3.8.2 Who Summarizes?

We can see a system such as Wikum used in a number of different scenarios. For instance, a single individual working to summarize a large discussion could derive benefit from some of the scaffolding and breaking down of summaries, as we saw in the Wikipedia field study. In a community setting, Wikum could be used by the small skilled groups of *moderators* already managing many discussion sites. These moderators currently focus on flagging and removing inappropriate content, and may well be interested in Wikum's alternative approach to curation. We saw in the Wikipedia study that collaborative summaries with a small number of people was promising but need some level of trust between editors.

Additionally, analogous to *social moderation* we envision contributions by a larger number of community members. After reading a deep thread, readers could summarize the content for future readers. Commentators could be required or encouraged to contribute short summaries of their comment (already common practice in some communities as a "TL;DR") or summarize a back-and-forth conversation in which they just participated. As argued above, only a moderate fraction of users' time need be spent on summarization in order to "keep up" with the arrival of new content. However, in the case of deliberative or polarizing discussions, we found that the Wikipedia frequent closers we asked about this were uniformly against the idea of having participants summarize discussions. They felt that participants could not be trusted to be unbiased in their work, and would simply try to get a leg up or always assume the other side was operating in bad faith.

If any user can edit or add a summary, more sophisticated tools for tracking, observing, and reverting changes are necessary. The current iteration of Wikum contains a wiki scoring and flagging mechanism similar to what is on Wikipedia. Other work has chosen to give commentators greater moderation power over the summaries of their own comments [189], but in our case they may be overly biased.

Crowd workers who have been tasked to summarize a discussion could also use this summarization workflow. As in the community case, we would need to build in robust spam filtering and verification, processes which have been explored in the literature [21].

3.8.3 Considering Community Values

Another interesting dimension that arose was how the design of a tool like Wikum can encode a community's principles. For instance, several participants in the Wikipedia field study rated highly the ability to hover over a comment author and see their edit count, account age, and user role, such as whether they were an admin. However, another participant pushed back, saying:

> "*...in an ideal world, I don't know if people are admins or what user length they have...I would want to weight everyone's opinion equally. I'm kind of worried that in a tool where it explicitly says [that]...that suddenly we're going to be weighing admin opinions more heavily.*" (P5).

Concerns were also raised in the interviews regarding the principle of consensus through deliberation as opposed to majority rule through voting, suggesting that tools to group and tally votes might also get some pushback, even if they would be useful to the closer. Future designs could bridge these sides through the use of strategic obfuscation; for instance, comment author information could simply display a warning if the edit counts or account age were unusually low in order to combat sock-puppeting, while keeping admin and other info hidden by default. Vote tallies could instead just present whether there was a supermajority or that votes were relatively close.

3.8.4 Production of a Public Summary Tree

Finally, while Wikum is a tool towards the *production* of a summary, the resulting output of a "summary tree," where a summary can expand into more summaries leading to original discussion, could serve as a navigable *presentation* of the closing statement as well. However, when we presented this as an option to our participants, none were interested in sharing their output with others and especially not the discussants. P4 of the Wikipedia field study stated:

> “*I’m not particularly keen on sharing it. Explicitly declaring that a certain editor’s argument had little weight is likely to be more contentious than making a more general statement about the type of argument.*”

P5 raised the point that with knowledge of Wikum, discussants could start demanding to see the output of Wikum in order to nitpick. Underscoring their point, P1 said:

> “*Oh no. Please don’t let [A] and [B] see that I summarized a thread thousand words long as ‘[A] and [B] continue to quibble without any new arguments.*”’

This reaction, though widely shared among our interviewees and study participants, is unfortunate as it is possible that a summary tree presentation could be helpful for newcomers to the discussion as a way to explore it or get an overview. However, building a public summary tree would likely require more work from closers to write additional diplomatic and comprehensible intermediate summaries beyond their existing closing statement.

3.9 Future Work

3.9.1 Summaries as Productive Outputs of Discussion

Our study also presents insights that could be valuable to systems and processes within peer production and deliberative communities beyond Wikipedia. Many platforms for discussion do not have definitive formal resolution processes like RfCs, focusing only on the deliberation aspect. For instance, in platforms like Kialo [171] or ConsiderIt [316], the discussion artifact, or resulting issue map, is the desired outcome. These platforms do not aim for a definitive end of the discussion but rather aim to have a fair and productive deliberation while mapping the space of opinions.

Whether or not the platform requires a definitive "task" to complete, systems seeking productive discourse might benefit from a more formal starting and ending nature of RfCs. Systems where discussions go on indefinitely or where threads with the same issue repeatedly arise might benefit from having a procedure that lets participants stop and move on to something else or work towards a conclusion. An interviewee mentioned:

> *"RfCs can bring even the most intractable disputes to a conclusion and allow editors to move forward despite holding extremely diverse opinions. A few times, I've even seen an entire topic area return entirely to quiet, 'normal' editing at the conclusion of a particularly important RfC"*

This emphasizes that RfCs provide a way for Wikipedians to move on and not get stuck on a particular issue. This is healthy for the community because editors can allocate their resources to different issues instead of wasting effort on a single one. This nature of RfCs may provide insights to platforms like Reddit's ChangeMyView, where there may exist participant fatigue around certain topics. Systems like Wikum [385] for collaborative summarization of discourse might be a vehicle for providing a sense of productivity or resolution. StackOverflow is similar in a way as redundant questions are frowned upon; instead, users are expected to update existing questions if possible.

3.9.2 Wikum for Summarizing Civic Discourse

A separate area where Wikum could be useful is towards the organization and summarization of civic discussions online. In recent years, several cities such as Madrid and Barcelona [12]. have pioneered online forums for citizen proposals, discussions, and voting, as a complement to in-person town halls that have limited participation. Individual proposals can rack up many hundreds of comments by citizens debating the merits and drawbacks and making suggestions. However, there are few government resources dedicated to actually reading the comments when a proposal is considered by the city council or put into action.

As part of an initial exploration into the feasibility of crowd summarization of these discussions, I conducted a pilot with study with 10 citizens of Madrid that are users of Decide Madrid, Madrid's citizen democracy platform. Users were instructed to use Wikum towards collaborative summarization of a discussion from a Decide Madrid proposal. Users were also given different interface treatments to consider how a collaborative platform like Wikum could bounce back from poor summaries, including the use of upvoting or flagging, in addition to wiki-like editing. Preliminary evidence suggests that flagging could be a useful way for participants to indicate when a summary is biased, poor quality, or missing content.

Interviews with participants were also encouraging, with all or almost all participants expressing interest in participating in summarization for this purpose and also strongly in favor of seeing citizen discussions actually used by city decision-makers. However, more study is needed.

3.9.3 Summarizing While Conversing

Wikum can be used to summarize a static discussion but does not currently support incorporating new comments. One interesting future line of work would be adapting Wikum to ongoing discussions. Thus a subthread that has been summarized may need to be updated when a user contributes a new comment to the discussion. Also, we could consider how people may want to "reply to" previously written summaries. It would be interesting to see whether the addition of summary tasks while conversing could potentially help with improving the quality of the ongoing discussion. For instance, participants could be prompted to reflect on each other's arguments or articulate what is shared knowledge before proceeding further.

3.9.4 Automatic Summarization

We incorporated automatic summarization techniques to help editors skim comments. There are other opportunities to incorporate machine learning. Techniques such aspect summarization of product reviews [207] could be repurposed towards grouping comments and providing default summaries of those groups to build upon. Users can also provide training data in a human-in-the-loop process to improve the quality of models. For instance, could machine learning help determine where to segment the discussion into discrete subparts? The data produced by this system could also be used to better build and train automatic summarization techniques for discussions.

This work also points to the need for better task definition and evaluation of discussion summarization. In our lab studies, we saw differences in summaries of the same content written by different people. Users also described different goals they were trying to achieve while writing a summary.

3.9.5 Authoring Tools that Expose Intermediate Work

Wikum is an example of an authoring tool and workflow that exposes intermediate states of production in the final presentation. In the case of a summary, the presentation of a summary tree can be useful for deeper exploration. There may be other artifacts that benefit from such an approach.

One example could be in the case of conducting exploratory data analysis via a tool like a Jupyter notebook. Research has shown that these authoring tools can produce messy final products requiring extensive scrolling around, little explanatory documentation [284], and lost test code that has been written over [170]. Instead of enforcing a linear document structure, users could be allowed to author multiple versions of a section of code via branching in the interface. Not only could users choose to hide or collapse parts of code [283], but they could grow a notebook from scratch using a hierarchical approach, either top-down or bottom-up.

Another case could be in the case of exploratory search interfaces [123], such as exploring online documents towards writing a report, or exploring through discussion forums for informational purposes. This could be particularly helpful when it comes to more intense information gathering tasks scattered in hard-to-gather places, such as thousands of redundant threads in TripAdvisor, or long single-thread discussions on CollegeConfidential, or health sites like PatientsLikeMe. Instead of losing all signals from the process of exploration, including rejecting certain trails, pursuing others, taking mental or jotted notes, this information could be externalized into a search interface that looks more like a workspace for a user to return to or for another person to pick up.

3.10 Conclusion

In this work, we designed, developed, and evaluated a workflow called recursive summarization for summarizing discussions and a system called Wikum that bridges discussion forums and wiki summaries. By bridging the two mediums of wiki and forum through embedding wiki summaries into a discussion structure at varying levels, we provide a process for editors to summarize portions of discussion and build upon each other's work. We also explore design decisions around an interface for readers to interactively explore a discussion, drilling deeper into a summary to get more information. From our lab evaluations, we found that editors created summaries productively using the Wikum interface and that the created embedded summaries were effective for helping readers get an overview of the discussion.

From a case study exploring summarizations of deliberations in Wikipedia, we found that frequent closers of deliberations must spend a great deal of time sifting through long discussions, rife with bickering and redundancy. This is partly why nearly a third of deliberations stay unresolved on the platform. From a field study of Wikum among Wikipedia editors, we find evidence that editors can use Wikum on their own to break down their work into manageable pieces and store intermediate state. We also have evidence that Wikum could facilitate collaborative summarization in Wikipedia.

Chapter 4

Tilda: Making Sense of Group Chat through Collaborative Tagging and Summarization

While group chat is becoming increasingly popular for team collaboration, these systems generate long streams of unstructured back-and-forth discussion that are difficult to comprehend. In this work, we investigate ways to enrich the representation of chat conversations, using techniques such as tagging and summarization, to enable users to better make sense of chat. This work was conducted in collaboration with Justin Cranshaw of Microsoft Research.

Through needfinding interviews with 15 active group chat users, who were shown mock-up alternative chat designs, we found the importance of structured representations, including signals such as discourse acts. We then developed Tilda, a prototype system that enables people to collaboratively enrich their chat conversation while conversing. From lab evaluations, we examined the ease of marking up chat using Tilda as well as the effectiveness of Tilda-enabled summaries for getting an overview. From a field deployment, we found that teams actively engaged with Tilda both for marking up their chat as well as catching up on chat.

4.1 Introduction

Group chat applications have seen considerable growth in recent years, especially for coordinating information work. By enabling quick, team-wide message exchange in different channels, these applications promise to minimize the frictions of group communication, particularly for distributed and remote teams. Many organizations use systems such as Slack [306], HipChat [141], Internet Relay Chat (IRC) [248], Microsoft Teams [231], and Google Hangouts Chat [110] to make decisions, answer questions, troubleshoot problems, coordinate activity, and socialize. As of 2016, Slack alone reported having over 4 million daily users [186].

However, chat systems can have a number of downsides. Unlike email or forums, chat is predominantly synchronous, with a heightened expectation for quick responses and a high volume of back-and-forth messages exchanged in rapid succession [35]. As a result, chat logs are often comprised of a great many short messages forming multiple distinct yet intertwined conversation threads, with little distinction made between messages that are important and those that are not. This can make it difficult for members of the group who are not present in the conversation in real-time to make sense of it after the fact—for example, when someone falls behind, goes on vacation, revisits old chat logs, or is a newcomer to the group. Perhaps because of this burden of sifting through chat conversations, users have criticized group chat as encouraging an overwhelming "always on" culture, and some organizations have chosen to cast it aside [149, 159].

To make group chat conversations more comprehensible, we can build off of sensemaking affordances designed for other textual domains, such as email, online forums [218], or documents and general information management [115]. For instance, *tags* can be added to important messages to contextualize them or differentiate them from unimportant messages, similar to labels in emails or highlighted sentences in long documents. Furthermore, adding *structure* to the conversation could allow related messages to be grouped, much like distinct threads in an email inbox. Finally, both of these affordances facilitate the *summarization* of long back-and-forth conversations into a condensed format, much like notetaking in live meetings. Although these approaches to sensemaking have been explored in asynchronous discussion [383, 388, 240, 385, 145], little work has explored how to enrich synchronous chat conversations, which has additional challenges.

In this work, we consider how to apply these techniques *in situ*, enabling participants to enrich their discussions *while they are conversing.* We explore a variety of ways chat participants can mark up portions of their chat to create enriched, structured representations that allow users to get a high level overview of a full conversation and to dive in to parts of interest. Furthermore, our approach does not require a dedicated notetaker, allowing our design to conform to the spontaneous nature of group chat discussions. We conduct our analysis through an iterative design process, beginning with needfinding interviews and design mock-ups, and culminating in lab studies and a field study of a prototype system.

From interviews, we learned about the information management practices of 15 active group chat users, finding that many interviewees have trouble keeping up with chat and often miss important messages while scrolling up to read through their backlog. To ground the interviews, we created mock-up illustrations of different synthesized representations of a chat conversation, each emphasizing different information extracted from the conversation and varying degrees of structure. Some designs made use of tags on individual messages, others focused on extraction of important quotes, while still others involved written abstractive summaries. From showing the designs to our interviewees, we found a preference for more structured designs as well as signals such as major *discourse acts* [291] in a conversation, where discourse acts are categories of statements that characterize their role in the discussion (e.g. “question” or “answer”).

Based on these findings, we developed Tilda, a prototype system built for Slack that allows discussion participants to collectively tag, group, link, and summarize chat messages in a variety of ways, such as by adding emoji reactions to messages or leaving written notes. Tilda then uses the markup left by participants to structure the chat stream into a skimmable summary view accessible within the chat interface.

The summaries become live artifacts that can be edited, referenced, and posted to particular channels and individuals. Users can dive in to points of interest by following links in a summary to its place in the original chat stream.

We evaluated Tilda through three studies. First, we performed a within-subjects experiment to measure the effort required for groups to mark their chat while executing a shared task. We compared Tilda to using Slack alone and using Slack with a shared online document for notetaking. From 18 participants, we found evidence that Tilda was the better tool for taking notes while participating in the conversation. In a second experiment, we used the artifacts created in the first study to investigate the effort for a newcomer to comprehend the past conversations. From 82 participants, we found that users looking over summaries and chat logs enriched by Tilda felt less hurried when catching up compared to the other conditions. Additionally, those who utilized the links within Tilda summaries to dive into specific chat messages had a lower mental load and performed better at finding information from the chat log while still taking less time overall. Finally, we conducted a week-long field study of Tilda within 4 active Slack teams of 16 users total, and observed that teams actively used Tilda to mark up content and also found Tilda to be effective for catching up or looking back at old content.

4.2 Related Work

Notetaking and Live Meeting Notes

A common technique for synthesis when it comes to synchronous conversations in particular is the practice of notetaking during meetings. Research has demonstrated that notetaking is beneficial both to individuals, in improving learning and comprehension [133, 174], and to teams and organizations, in improving knowledge management practices and fostering collaboration [211]. During live meetings, it is common for teams and organizations to assign someone the role of designated notetaker [86], who may find it difficult to participate in the conversation due to the cognitive effort and split attention required to take notes [373, 265, 263]. Summary writing, although different from notetaking in that its primary source is an existing text, exhibits similar cognitive burdens, dividing the summary writer's attention between the act of reading and comprehending and that of writing the summary [176]. Due to the cognitive load of synthesizing conversation, we consider how more lightweight techniques such as tagging or inline notes in the chat could make notetaking easier. We also consider how the work could be broken down and distributed among participants, both to lower individual load and spread the benefits of participation.

Conversational User Experiences

In order to integrate seamlessly into chat conversations as they are ongoing, our Tilda prototype is developed as a Slack bot [202], exposing its functionality to the participants within their conversation. Chatbots have a long history in research [296], from initial explorations for fun and entertainment [355], to modern assistants offering a conversational interface to complex tasks [27, 36, 52, 85, 335]. Our system differs from many of these bots, in that it does not rely on natural language understanding [297], and is instead command driven, reacting only to specific user-input commands and actions. Several command-driven chatbots initially gained popularity in IRC communities [30], including Debian MeetBot [65], which is still actively used by organizations such as Ubuntu and Wikimedia to take notes during IRC meetings, or Zakim [349], which is in use at the W3C. MeetBot allows the leader of a chat meeting to designate the start and end of the meeting and enables participants to add different forms of notes to a running list of notes using hashtag commands. Similarly, Zakim is used during meetings for setting agenda items, reminders, speaking queues, and meeting scribes. While inspired by MeetBot, our prototype tool does not require scheduled meetings but can be used for more informal group chat conversations, with topics shifting continuously and people coming in and out throughout the day.

4.3 Needfinding Interviews for Making Sense of Group Chat

We began by interviewing active group chat users to understand how, why, and how often they go through prior chat conversations, and their strategies for and frustrations with making sense of long streams of chat messages.

4.3.1 Methodology

We conducted semi-structured interviews with 15 people who use group chat on a daily basis (6 female, 9 male, average age of 30.0). Interviewees were recruited through social media postings, email lists, and word-of-mouth, and were compensated $20 for their time. Individuals came from a diverse set of group chat teams, including tech companies, research groups, communities of interest, and co-working spaces. Groups ranged from as small as 4 people to over 500 people and from exchanging a few messages a day to thousands. Interviewees used a multitude of applications for group chat, including 11 on Slack, 4 on Microsoft Teams, 1 on HipChat, and 1 on WeChat.

We began by asking interviewees to open up the chat application for their most active chat group. We asked about how interviewees access their chat, their frustrations with using group chat, and their practices for managing the volume of chat messages they receive. We next sought to understand what content interviewees found important within their chat and which signals determine that importance. We asked interviewees to find an important conversation in their chat of which they were not a part and explain how they determined it was important and what they wished to glean from it. We then presented mock-up designs showing four different synthesized versions of the same conversation to them in randomized order, to probe their opinions about the type of information shown and the presentation of that information.

Interviews were conducted remotely by the first author and lasted 45-90 minutes. They were recorded and then transcribed using a paid transcription service. Then, the first author went through the transcripts and coded them for themes using an open coding approach [38]. Through multiple iterations along with periodic discussions with the rest of the research team, the coding led to 71 codes, from which the following major themes were selected. Because of the low number of interviewees, our interview findings should be regarded as indicative.

4.3.2 Current Experiences with and Strategies for Managing Group Chat

Participants have an "Always On" Mentality but Still Fall Behind

Almost all (14/15) interviewees kept their group chat application open on their computer or phone the entire day, echoing reports that users of Slack have it open 10 hours on average per weekday [157]. Interviewees cited many reasons for being continually present, including being "on call" to answer urgent messages, seeking to gain an ambient awareness of others' activities, a concern about "missing out", and disliking having to deal with a backlog of missed conversations. But several interviewees acknowledged downsides of continually being on chat, with one saying:

> "*I think there's a lot of content that I don't need to consume. I've read [that] content switching is distracting and bad for productivity...But I hate having unread notifications.*"

Most interviewees (11/15) also mentioned checking chat while not working or on vacation, and checking it more often than they would have liked. Despite their efforts, falling behind was a common occurrence (13/15 interviewees). Some interviewees blamed the volume of messages while others had trouble distinguishing relevant information:

> "*There are so many things happening at the same time...I had a very hard time [determining] what are relevant for me, and what are the things I don't really need to care about at all.*"

Still others purposefully let messages go unread in certain channels or groups because the ratio of important to unimportant messages was low or they had only a passing interest in the topic.

Newcomers are Overwhelmed by Chat History

Besides active members, newcomers are another population that may desire to go through concluded conversations. A few interviewees (4/15) talked specifically about the newcomer experience of joining a chat group. They described it as overwhelming and tedious, but they still felt compelled to go back over the chat history to get better acquainted with the team and the work. For instance, one interviewee said:

> "*...there was a whole history of stuff that I wanted to know about so that we could not reinvent the wheel, so that we could understand where ideas are coming from...It was not so much about missing stuff. It was more coming into a new thing...wanting to know what is it? Because you just can't read back through it all.*"

Strategies for Catching Up are Unsatisfactory

When looking back through chat history, either to catch up or to get acquainted with a group, we found that the dominant strategy (9/15) was to skim content by scrolling up in their chat window. However, several expressed frustration with this strategy, with one interviewee saying:

> "*Scrolling is basically the big issue, which is that you've got this giant timeline of stuff...You can only scroll and skim back through so many views on the viewport before you start getting tired of looking.*"

Other interviewees echoed this sentiment, pointing to how chat logs are poorly segmented by discussion topic, contain a great deal of back-and-forth before reaching a conclusion, and intersperse important information with humor or chit-chat, providing little ability to distinguish the two. One interviewee said:

> "*...there's a lot of conversation, and it all concludes with some result...all I want is results...then I wouldn't have to read 300 back-and-forths.*"

When falling behind, several interviewees (6/15) also simply chose to ignore missed messages, assuming they were irrelevant by then or that important information would reach them eventually, such as by email. This strategy exacerbated issues such as questions that were continually re-asked, or important requests that went unanswered. One interviewee said:

> "*[Someone] was requesting help for something...I knew when I read it that everyone was going to ignore it because it was going to get lost in the Slack channel...it was a really important thing but it was just a lot easier to ignore...it just sort of gets pushed up...*"

Even though interviewees felt that important information would eventually reach them, several (5/15) could remember specific instances when they had missed important information that they wished they had known about. In these cases, someone neglected to mention them in the conversation, or an announcement was made that got lost among other messages, or they had a passing interest in a channel but no way of occasionally dipping in to catch up on important happenings.

Recalling or Re-finding Information is Hard

Another way to explore a long chat stream is to use search to filter for specific conversations. Half of the interviewees (7/15) had trouble searching back over chat conversations to find information. Interviewees, when trying to recall conversations they were a part of, needed to remember particular phrases that were said or other details, with one saying:

> "*...if you don't know exactly what you're looking for, or if you misremembered the word...search begins to be fairly limited...Usually you'd need two to three bits of information. A word, a person...[otherwise] there might be months' worth of stuff...*"

Interviewees who couldn't pinpoint information with search resorted to scrolling in the surrounding area of the search results, encountered the same issues with scrolling mentioned earlier.

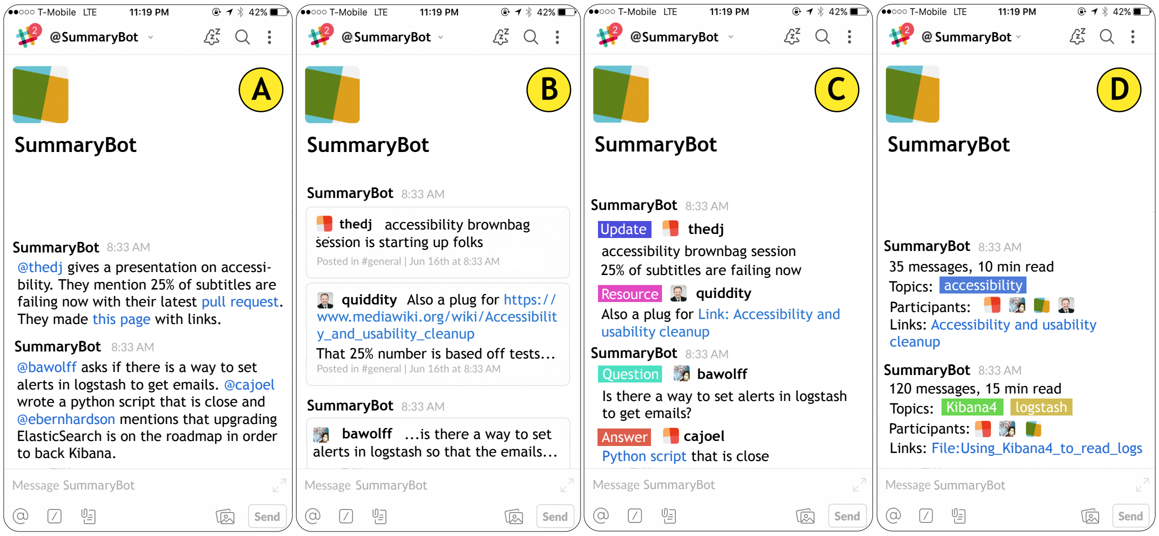
Related to the strategy of expecting important information to arrive through multiple avenues, a few interviewees (4/15) also described conversations spilling over from chat into email, making it harder to retrace what happened. One interviewee said:

> "*It's especially annoying if this conversation started here and then there was an email thread, and it was hard to interlace the two chronologically.*"

Another interviewee, catching up from vacation, made a note to respond to an unanswered request in chat but missed that it had been responded to in an email. Thus, using multiple channels for pushing out information may make it difficult to recall where conversations took place.

Existing Processes for Organizing Information are Cumbersome

In response to difficulties with finding or catching up with chat conversations, some interviewees described policies the group had instated to collect knowledge. However, many of these were unsuccessful because of the cumbersome nature of the process, leading to lack of adherence to the policy or lack of maintenance over time. For instance, several interviewees (5/15) had a separate knowledge store, such as a community Q&A site, collaborative documents, or a wiki. One interviewee, discussing finding answers to questions, said he preferred to search the chat history instead of his company's internal community Q&A site because people often failed to post to the Q&A site or update their post with the answer. This was considered a documentation chore, uncoupled to the main goal of getting the question answered, despite being considered a good practice in the team. Two interviewees also mentioned how people summarized accumulated pinned messages in Slack into Google Docs files; however, the files were rarely used and quickly forgotten due to their lack of visibility in the chat system. Another interviewee complained about how it always fell to the same people to organize information from chat, highlighting the diffusion of responsibility due to the group setting.



SummaryBot

SummaryBot 8:33 AM
@thedj gives a presentation on accessibility. They mention 25% of subtitles are failing now with their latest pull request. They made this page with links.

SummaryBot 8:33 AM
@bawolff asks if there is a way to set alerts in logstash to get emails. @cajoel wrote a python script that is close and @ebernhardson mentions that upgrading ElasticSearch is on the roadmap in order to back Kibana.

SummaryBot

SummaryBot 8:33 AM
thedj accessibility brownbag session is starting up folks
Posted in #general | Jun 16th at 8:33 AM

quiddity Also a plug for https://www.mediawiki.org/wiki/Accessibility_and_usability_cleanup
That 25% number is based off tests…
Posted in #general | Jun 16th at 8:33 AM

SummaryBot 8:33 AM
bawolff …is there a way to set alerts in logstash so that the emails…

SummaryBot

SummaryBot 8:33 AM
Update thedj
accessibility brownbag session
25% of subtitles are failing now
Resource quiddity
Also a plug for Link: Accessibility and usability cleanup

SummaryBot 8:33 AM
Question bawolff
Is there a way to set alerts in logstash to get emails?
Answer cajoel
Python script that is close

SummaryBot

SummaryBot 8:33 AM
35 messages, 10 min read
Topics: accessibility
Participants:
Links: Accessibility and usability cleanup

SummaryBot 8:33 AM
120 messages, 15 min read
Topics: Kibana4 logstash
Participants:
Links: File:Using_Kibana4_to_read_logs

Figure 4-1: Some examples of mock-ups shown to interviewees to compare and contrast different synthesized chat designs: A) abstractive, B) extractive, C) discourse act labels, D) high level signals.

Summary

We found that many interviewees spend a significant amount of time scrolling through their chat history, despite being continuously available on chat, and face frustrations with differentiating content when doing so, leading to missed important information. We also saw how conversations that start in chat sometimes get picked up in email or vice versa, making them hard to follow and re-find. This suggests that tools could better bridge and link more synchronous communication systems such as chat to more asynchronous ones such as email. Similarly, we saw that attempts to synthesize information from chat failed because they were poorly integrated, due to being in a separate location and with a workflow separate from chatting. This suggests that tools for enriching or synthesizing chat should be tightly integrated into the chat environment, and any artifacts created should also be coupled to the original discussion.

4.3.3 Preferences for the Content and Presentation of Synthesized Chat Designs

Next, we sought to learn from our interviewees what information from a chat conversation is useful for determining importance, as well as what presentation of that information is best for gaining an overview quickly. We did this by asking interviewees to find an important chat conversation from their chat history to talk about as well as give their impression of four different design mock-ups that we prepared beforehand. We presented the design mock-ups to interviewees in a randomized order, and for each, asked interviewees what aspects they liked and disliked. At the end, we asked interviewees to compare the designs and discuss which ones they preferred and why.

The mock-ups were conceived by surveying existing applications for enriching or synthesizing conversations. They were also chosen to encompass a diversity of types of information, from excerpts to topics to discourse acts, as well as a range of presentations, from less structured to more structured, to elicit interviewees' reactions. Figure 4-1 shows examples of the four mock-up types. Design A presents a written abstractive summary of the discussion in the form of short sentences, inspired by the practice of notetaking in meetings. Design B is an extractive summary made up of important excerpts taken directly from the chat log, inspired by tools like Digest.AI [68] or Slack's Highlights feature [307]. Design C augments excerpts of the conversation by tagging them with major discourse acts, similar to tools like Debian MeetBot [65]. Finally, Design D showcases high level signals, such as main participants, number of messages, topic tags, and a subject line, inspired by affordances in major email clients. We created two examples for each design, with conversations taken from the same chat from a Wikipedia IRC chat log. We asked interviewees to assume that all designs are manually created to sidestep concerns about perceived feasibility of automation.

A Purely Extractive Approach Lacks Context

Only one interviewee preferred a purely extractive approach (Figure 4-1-B) for getting an overview, stating that she preferred to read people's contributions in their own voice. However, most interviewees did not like this design because of the loss of context, with one interviewee stating:

> "*A lot of these messages are very much conversational, and so unlike an email where everything is self contained, it's a flow. So just pulling out a single message does lose some of that important context.*"

This was surprising given the number of existing tools that use an extractive approach. Two interviewees were aware of the Slack Highlights feature [307] that shows automatically extracted important messages, but expressed the same concern.

A Purely Abstractive Approach Lacks Structure

Alternatively, only 3/15 interviewees liked the purely abstractive approach (Figure 4-1-A). This was also surprising given that abstractive summaries of a conversation would likely be the most labor-intensive to create and is often considered a gold standard in summarization tasks. The interviewees that liked this design liked that it was possible to gain a comprehensive understanding of what happened, while other designs offered an indication but would need further investigation. However, most interviewees objected to this design because they found it too difficult to skim due to the lack of structure. One interviewee said:

> "*I have no way of knowing almost until I finished this thing whether or not I'm interested. It doesn't save me any time triaging.*"

Two interviewees also mentioned needing to trust the writer of the summary and were concerned about variability in quality.

Signals about Topic, People, and Volume are Informative and Easy to Skim

Eight interviewees liked the design exploring different high-level signals about a conversation (Figure 4-1-D), with most commenting on the additional structure provided.

One interviewee said:

> "*I can decide on the outset if I care about the thing that was discussed or not, and if I don't care, then I move on. I don't like the clutter of having long or multiple messages.*"

Many interviewees found signals such as topic keywords, a main subject line, major participants, and the number of messages or an estimate of reading time to be informative.

Discourse Act Tags Add Context to Extracted Messages

Finally, the design exploring the use of major discourse acts as labels to group notes was by far the most popular, with 14/15 interviewees preferring this design (Figure 4-1-C). Given the additional structure, interviewees felt they had a greater ability to skim and home in on specific categories of interest, such as unanswered questions, which was difficult in the abstractive or extractive designs. But unlike the design with only high-level signals, this design still provided information about what occurred in the discussion. One interviewee said:

> "*I love the tags. I love the fact that sometimes you have a question and now the question leads to an answer...It tells me how to read the content.*"

The improvement over a purely extractive approach was the ability for the discourse acts and links between them to provide a narrative for the extracted messages.

Given the emphasis that interviewees placed on major actions over the course of a conversation, we asked interviewees to consider what kinds of discourse acts they would want to have highlighted. The following discourse types were mentioned:

- **Action items**: Several interviewees mentioned wanting a way to track assigned action items or any follow-up to-dos that resulted from any kind of discussion.

- **Troubleshooting**: Several interviewees also mentioned the importance of marking problem statements, the resolution of troubleshooting discussions, as well as suggestions or ideas to solve them. Interviewees also wanted to easily see which problems were still ongoing.

- **Deliberation**: Interviewees mentioned having many scheduling discussions or debates. They thought of labeling these with a problem statement along with a decision marking the outcome or pros and cons labeled separately.
- **Questions and answers**: Similarly to problems and solutions, interviewees wanted to highlight questions, along with their answers, as well as any unanswered questions.
- **Announcements, links, tips**: Finally, interviewees saw a use case for labeling announcements and links to items, as well as observations, tips, or other useful one-off information.

Hierarchical Exploration Manages Volume and Provides Agency

Finally, interviewees described how they would prefer to interact with synthesized representations of chat. Some interviewees (4/15) desired some sort of ability to explore hierarchically, whether that be from the summary to the original discussion or from a shorter summary with high-level signals, to a longer summary that contained excerpts. One interviewee stressed the importance of controlling exploration, saying:

> "*I want to scroll through it and zoom in and out of it...skim, but skim with a bit more intent. I might be more likely to use...something a bit more interactive. I don't want to just be told...I want to be helped.*"

Another interviewee wanted a different level of depth depending on how much conversation they had missed; the more they missed, the shorter each individual summary should be.

Summary

From the feedback that the mock-ups prompted, we found that interviewees preferred a high degree of structure to aid their sensemaking. At the same time, they were interested in cues that could provide context about what happened in the discussion. This feedback suggests that a hybrid approach combining structured high-level signals about a conversation with important excerpts marked with their discourse act could be both easily skimmable yet contextual. Finally, we found that interviewees desired the ability to use summaries to guide deeper exploration. This suggests that summary views could have different hierarchies of synthesis, with a shorter initial representation leading to a longer one, eventually leading to the original discussion.

4.4 Tilda: A Tool for Collaborative Sensemaking of Chat

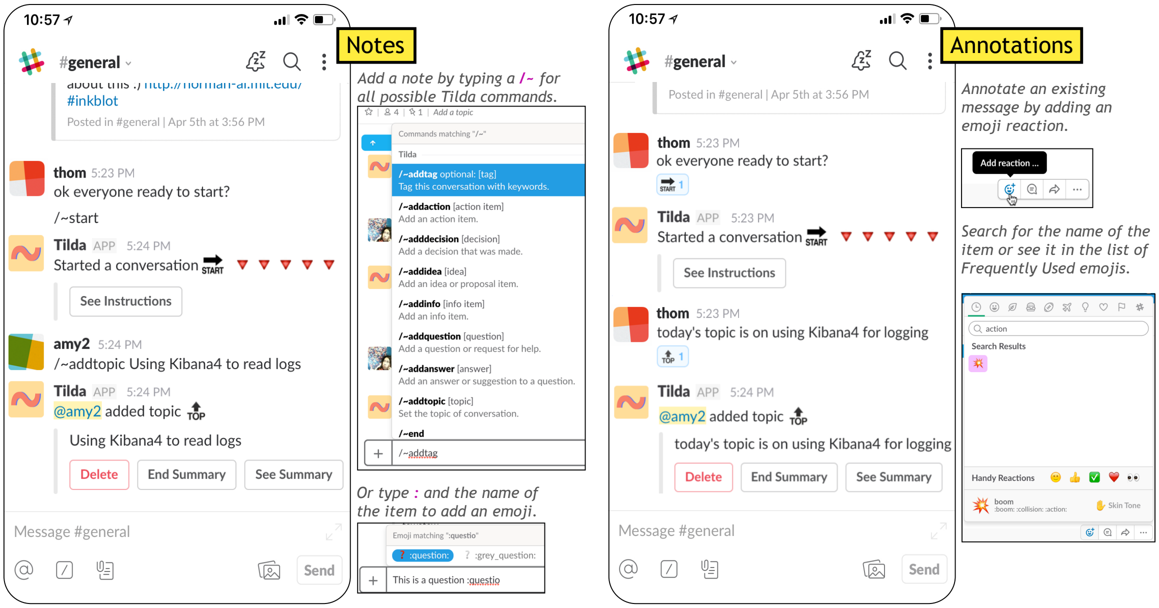
Building on the findings of our interviews, we developed a prototype system called Tilda[1], instantiated as a Slack application, for participants in a group chat conversation to collectively mark up their chat to create structured summaries, using lightweight affordances within Slack.

4.4.1 Enriching Chat Conversations using Notes and Tags

Techniques for Enriching Chat

Tilda provides two main techniques for enriching a chat conversation, as shown in Figure 4-2. The first way is through inserting a **note** while in the course of conversation. A user may add a note by using a custom *slash command*, Slack's feature for invoking commands within the dialog box, or by adding a custom *inline emoji* to the text of their message. Slash commands allow users to type a slash in order to pull up an auto-completed list of commands. For this reason, all Tilda commands are prepended with a tilde. Some types of notes consist solely of the command, such as a note to designate the start or end a conversation. Other notes contain textual content, such as the marking of a conversation's topic or the addition of a question.

[1] Visit `tildachat.com`. Tilda sounds somewhat like pronouncing "TL;DR" (too long; didn't read). The logo for Tilda is a tilde.



about this :) http://norman-ai.mit.edu/ #inkblot
Posted in #general | Apr 5th at 3:56 PM

thom 5:23 PM
ok everyone ready to start?
/~start

Tilda APP 5:24 PM
Started a conversation START ▼ ▼ ▼ ▼ ▼
See Instructions

amy2 5:24 PM
/~addtopic Using Kibana4 to read logs

Tilda APP 5:24 PM
@amy2 added topic TOP
Using Kibana4 to read logs
Delete | End Summary | See Summary

Add a note by typing a /~ for all possible Tilda commands.

4 | 1 | Add a topic
Commands matching "/~"
Tilda
/~addtag optional: [tag]
Tag this conversation with keywords.
/~addaction [action item]
Add an action item.
/~adddecision [decision]
Add a decision that was made.
/~addidea [idea]
Add an idea or proposal item.
/~addinfo [info item]
Add an info item.
/~addquestion [question]
Add a question or request for help.
/~addanswer [answer]
Add an answer or suggestion to a question.
/~addtopic [topic]
Set the topic of conversation.
/~end
/~addtag

Or type : and the name of the item to add an emoji.

Emoji matching ":questio"
:question: :grey_question:
This is a question :questio

Posted in #general | Apr 5th at 3:56 PM

thom 5:23 PM
ok everyone ready to start?
START 1

Tilda APP 5:23 PM
Started a conversation START ▼ ▼ ▼ ▼ ▼
See Instructions

thom 5:23 PM
today's topic is on using Kibana4 for logging
TOP 1

Tilda APP 5:24 PM
@amy2 added topic TOP
today's topic is on using Kibana4 for logging
Delete | End Summary | See Summary

Annotate an existing message by adding an emoji reaction.

Add reaction ...

Search for the name of the item or see it in the list of Frequently Used emojis.

action
Search Results
Handy Reactions
boom
:boom: :collision: :action:
Skin Tone

Figure 4-2: The main techniques for adding metadata in Tilda include notes and tags. On the left, the chat is enriched in real time by injecting notes using slash commands or inline emojis. On the right, the chat is marked up by adding tags to pre-existing messages using emoji reactions.

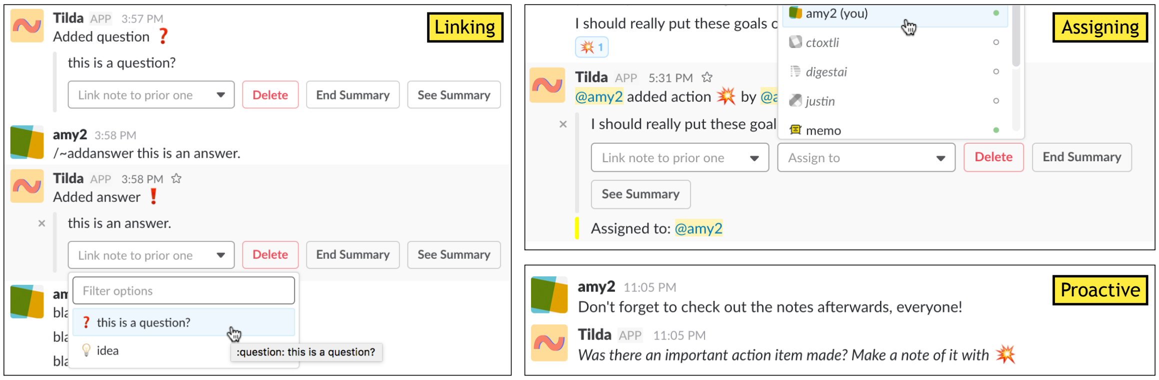
Each note gets added as a chat message to the transcript of the chat log when they are created.

The second way is through **tagging** of existing chat messages using custom *emoji reactions*, a feature in Slack, as well as common in other messaging systems such as Facebook Messenger, where any user can attach an emoji to the bottom of an existing message. Users can use this method to tag any pre-existing message going back in time, and so can choose to mark up an old conversation or one as it is ongoing. Users can use tags to designate messages as the start or end of a conversation or mark messages with their discourse act, such as a question or an answer. Unlike slash commands and inline emojis, one can add an emoji reaction to anyone's chat message, not just their own.

For each item added, whether by note or tag, the Tilda application posts a message in the chat documenting the action and allows the user to undo their action, toggle to see the current state of the items in the conversation, or interact with the items in other ways.

Label	Command	Emoji	Function
Action	\~addaction	💥	Add action item
Answer	\~addanswer	❗	Add answer item
Decision	\~adddecision	🤝	Add decision item
Idea	\~addidea	💡	Add idea item
Question	\~addquestion	❓	Add question item
Topic	\~addtopic	🔝	Add topic of conversation
Tag	\~addtag		Add custom tag to conversation
Start	\~start	➡️	Start a new conversation
End	\~end	🔚	End current conversation

Table 4.1: List of discourse act items and their commands and emojis, as well commands and emojis related to conversation-level markup, including adding a topic or custom tag and starting or ending the conversation.



Tilda APP 3:57 PM
Added question ❓
this is a question?

amy2 3:58 PM
/~addanswer this is an answer.

Tilda APP 3:58 PM
Added answer ❗
this is an answer.

❓ this is a question?
💡 idea
:question: this is a question?

I should really put these goals c
💥 1

Tilda APP 5:31 PM
@amy2 added action 💥 by @a
I should really put these goal
amy2 (you)
ctoxtli
digestai
justin
memo
Assigned to: @amy2

amy2 11:05 PM
Don't forget to check out the notes afterwards, everyone!

Tilda APP 11:05 PM
Was there an important action item made? Make a note of it with 💥

Figure 4-3: Examples of linking a Tilda item to a prior item, assigning an Action item to a member, and getting a proactive nudge to annotate a message.

Categories of Tags or Notes

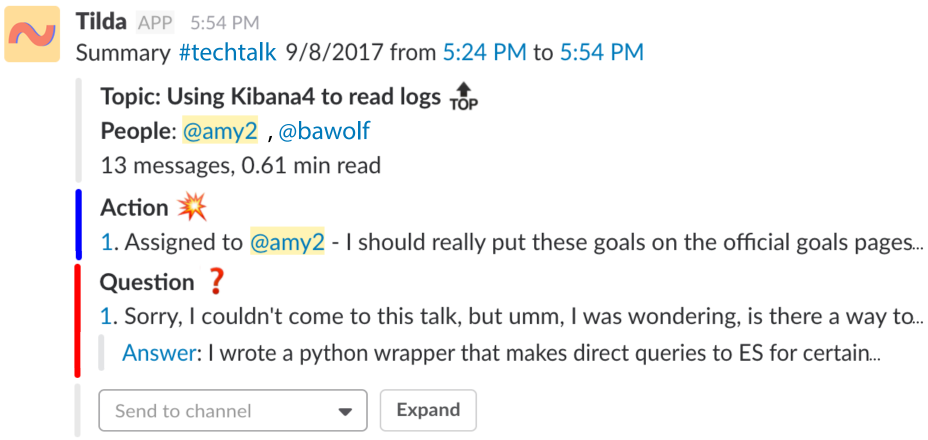
Using either of these two techniques, users can add a variety of metadata to their chat conversation (see Table 4.1 for a complete list). First, as mentioned above, users can mark the beginning and end of conversations as a way to segment the chat stream and **group** a series of items together. This can be done using either the note or tag technique. For convenience, conversations also automatically start whenever a new piece of metadata is added to the chat, and they automatically end if there is no activity for 20 minutes, though this can be undone if it was premature. In between start and end markers, users can mark up the chat by contributing Tilda items to an ongoing summary of a conversation. The possible discourse acts, as seen in the first five rows in Table 4.1, correspond to the types of discussion actions that interviewees wished to have highlighted. In addition, users can add a topic sentence to a conversation or add a custom topic tag to the conversation, two signals our interviewees found informative.

Adding Additional Context

In addition, we provided other abilities to add structure based on findings from our interviews. First, a user can **link** a Tilda item to a prior one, as shown in Figure 4-3. This can be used when an item should be seen in context with another item for it to be better understood. For instance, an Answer item could be linked to its corresponding Question item. Linking is facilitated by a dropdown menu in the chat message that the Tilda application posts. For Action items in particular, users can also **assign** the item to a person who is a member of the channel. This was added because several interviewees were interested in tracking to-dos that arose due to discussion. Any user can assign the Action item or re-assign it at a later point in time.

Encouraging Participation

Finally, to encourage or remind users about notetaking, Tilda proactively posts suggestions to add a tag when it notices certain activities, as seen in Figure 4-3. These activities were determined manually and encoded in Tilda as explicit rules. For instance, if a user *stars* a recent message, a feature in Slack to private pin messages to a list, Tilda will post a suggestion to annotate it with a discourse act. Second, we manually devised a number of phrases associated with each discourse act type based off of conversations we saw in pilots, such as “remember to” with “Action”. When Tilda sees such a phrase, it posts a suggestion to annotate the message with the corresponding discourse act. In the future with more data, one could imagine moving to machine learned suggestions trained by prior tagged messages.



Tilda APP 5:54 PM

Summary #techtalk 9/8/2017 from 5:24 PM to 5:54 PM

Topic: Using Kibana4 to read logs 🔝

People: @amy2 , @bawolf

13 messages, 0.61 min read

Action 💥

1. Assigned to @amy2 - I should really put these goals on the official goals pages...

Question ❓

1. Sorry, I couldn't come to this talk, but umm, I was wondering, is there a way to...

> Answer: I wrote a python wrapper that makes direct queries to ES for certain...

Figure 4-4: Example Tilda summary generated from user tags and notes. The summary is grouped by discourse act, expandable, and each note is linked to its place in the original chat.

4.4.2 Synthesizing Chat Conversations using Structured Summaries

The notes and tags that users leave behind using Tilda can be immediately used by readers scrolling up through the chat log. Tilda also gathers them into structured summaries that allow a reader to get an overview of a discussion as well as dive in to the original chat.

Presentation of Summaries

Figure 4-4 shows an example of a summary in Tilda. Based off of feedback from interviewees, the summary includes signals about the conversation, such as number of messages and estimated read time, major participants, any custom tags that users have added to the conversation, and a topic sentence if it exists. It also presents the items that users added grouped and colored by their discourse act type. If an item was linked to another item, it appears underneath and indented to the right. Because users may leave many items in a single conversation, we only show a subset in the summary with the ability to expand to see all. The subset is determined using a heuristic that prioritizes categories like Topic and Action and limits each category to the first two items left chronologically. Upon expanding, users may sort all the items chronologically or grouped by category. Each item is also preceded by a hyperlink pointing to where it originally took place in the chat log, providing the hierarchical exploration and deep integration between summary and discussion that interviewees desired. In addition, because all items in the summary originate as markup in the original chat log, any edits to the content or markup in the original discussion automatically updates the summary, making it a live artifact wherever it is displayed.

Delivery of Summaries

One way that summaries can be delivered is through **following** the summaries of a particular channel. Any user can, in another public or private channel, set that space to follow the summaries of a public channel using the slash command `\~followchannel #channelname`. Users can specify parameters in the `followchannel` command to limit summaries to only those containing a particular participant or tag. From then on, all summaries matching the parameters and generated in the original channel will get posted to its designated places. In this way, Tilda could be used to take discussions from a smaller or private channel and have them summarized to a larger or public one.

One potential way to set up Tilda is to create team-wide "summary channels" that follow the summaries of one or more other channels. Another more personalized way to use Tilda is for a user to subscribe to the summaries from one or more channels in their direct message with Tilda. Finally, users also have the ability to selectively send a single summary to a channel using a dropdown, as seen in Figure 4-4.

4.4.3 System Implementation and Considerations

Tilda is implemented as a Slack application, with messages from Tilda arriving in the chat log, similarly to a chatbot. It is built on top of the Microsoft Bot Framework, an SDK that allows one to develop chatbots for a number of applications at once, and the Slack API. The backend server is built in Node.js and interfaces with a MongoDB database.

Several considerations went into the implementation of Tilda. First, we chose to develop a Slack application over developing a separate chat system or a browser extension because Slack applications can be quickly installed to any team already on Slack using OAuth authentication. Additionally, users can use it in any browser of their choosing on mobile, tablet, or desktop. We chose to implement for Slack over other chat systems such as IRC because of the ability to use Slack-specific features such as custom slash commands and emoji reactions, as well as create interactive and dynamic prompts within chat messages. Finally, we chose to build sensemaking capabilities into a chat system as opposed to designing a separate system that imports chat messages. We chose this direction after encountering difficulties with understanding chat after the fact, which we uncovered while piloting interfaces and workflows for marking up a pre-existing chat log.

However, these decisions also required us to make some trade-offs due to the limitations of Slack's API. For instance, the only way to communicate with users or add affordances beyond commands and emojis is to post a message in the chat as a bot. But due to the space they take up, messages posted by Tilda could pollute the chat stream. Additionally, summaries can only be presented via a chat message, which may be difficult for users already juggling multiple channels. A more integrated approach might have summaries overlaid on top of or directly alongside the original discussion. In the future, these ideas could be explored in a novel chat system or an extension that can alter the existing interface. In the meantime, our prototype allows us to quickly experiment with and deploy techniques for enriching and synthesizing chat in real-world settings.

4.5 Evaluation

We conducted two lab evaluations of Tilda to study how easy it is to enrich chat conversations while chatting as well as to study the experience of catching up on missed conversations using structured summaries. While these lab studies enabled us to examine specific facets of Tilda usage in detail, they were necessarily conducted under artificially constrained setting. To examine Tilda in more naturalistic chat settings, we also conducted a field study, where we observed expert Slack users from real organization use Tilda while they conducted their normal activities.

4.5.1 Study 1: Marking Up Chat Conversations While Chatting

In the first lab study, we considered the common scenario where chat participants wish to make note of important discussion items while they also actively conversing. We conducted a within-subjects experiment that compared using Tilda for keeping notes to more traditional methods such as collaborating on a shared online document for notes, or not taking notes at all.

While it is common for group chat conversations in real organizations to be partially asynchronous, focusing on notetaking during *active* discussions enabled us to explore the cognitive load and cost of switching contexts between participating in chat and marking content with Tilda, as it compares to using an online collaborative document. We were also interested in understanding whether any benefits from notetaking would justify the added overhead of keeping notes.

We recruited 18 participants (mean age 36.6, 8 female, 10 male) from UpWork, a paid freelancing platform, at the rate of $20 per hour, with each participant working around 2.5 to 3 hours in total depending on their assigned conditions. Participants were all based out of the U.S., native or fluent English speakers, and somewhat or very tech-savvy, though 6 participants were new to Slack. Participants were placed randomly into groups of 3, with 6 groups total.

Discussion Tasks

We devised two collaborative tasks that each group would perform together. The tasks were chosen because they were comprised of many smaller parts that needed to fit together, and they involved deliberation as opposed to simply compiling or coordinating information. The tasks were:

- **Story**: Collectively come up with a new T.V. show based on the show Friends. Participants were asked to come up with the cast, location, and the plot of a 5-episode season.
- **Travel**: Plan a month-long cross-country roadtrip in the U.S. Participants were asked to pick 5 major cities and national parks and other landmarks to visit, as well as the route, transportation, and accommodations.

Experiment Design

Every group of 3 completed the Story task first, followed by the Travel task. Each task was completed in one of the following three conditions:

- **Tilda**, where the group used Slack with Tilda to discuss the task and mark up their chat,
- **Doc**, where the group used Slack to discuss the task and take notes using a shared Google Doc, and
- **None**, where the group used Slack to only discuss the task.

Since there were two tasks per group, each group participated in a pair of conditions. Thus, for every pair of conditions, two groups out of the six groups total were assigned that pair. To account for ordering effects, we counterbalanced the condition order, so groups with the same pair of conditions received a different condition first.

To start a study session, we invited everyone in a group to a Slack channel, where we spent 30 minutes on an icebreaker and a tutorial on Slack administered via a Word document shared with the group. Then users worked on their first task for 45 minutes and completed a post-task survey rating their experience. They were then invited to a different Slack channel where they worked on the second task for 45 minutes, completed the same survey, and then completed a survey comparing the two conditions. They then collectively participated in a debriefing discussion in Slack with the authors about their experience where we asked them to compare conditions.

Before the TILDA condition, we gave users a 30 minute tutorial covering advanced Slack features and Tilda, again using a Word document shared with the group. During this session, users got acquainted with Slack slash commands, inline emojis, and emoji reactions, as most of our subjects did not have much familiarity with these features. In addition, the tutorial provided a basic overview of Tilda, covering the different types of notes and tags one could leave using Tilda. Before the DOC condition, we gave users access to a shared Google Doc for notetaking. There was no tutorial for Google Docs as all our users stated they were experienced Google Docs users.

In the DOC and TILDA conditions, we required users to keep track of their conversation using the provided tools. Users were also told before the study that they would debrief the authors afterwards about what they decided so as to motivate them to keep better notes.

Results

We compare each condition against each other. Due to the small sample size, the results are not statistically significant, except where indicated otherwise. Instead, we present more qualitative findings and observations that should be regarded as indicative.

TILDA versus DOC: All 6 users that were in both TILDA and DOC conditions marked Tilda as substantially better at keeping track of what happened in the discussion. Additionally, 4/6 users thought Tilda was somewhat or a lot better for participating in the discussion, and most preferred to use Tilda for the same task again (5/6). One user said:

> "*Honestly now that I know about Tilda I would never use Google Docs for brainstorming ideas with others. Tilda is way simpler.*"

Other users talked about being more organized with Tilda:

> "*...the Google Doc was hard to follow if you didn't know what it was already about but I feel Tilda kept all of our ideas organized and made it easier to follow.*"

5/6 users marking that Tilda was a lot better for looking back over the discussion. However, 3/6 users found Google Docs to be easier to use for notetaking than Tilda. This may partially be because they just learned Tilda but were experienced Google Docs users:

> "*I think because I use Google Docs regularly, it makes more sense to me. But Tilda captures a conversation better.*"

We also analyzed post-task survey ratings of all TILDA conditions and all DOC conditions, finding that people in TILDA conditions rated themselves on a 5-pt Likert scale as more successful in accomplishing their task (N=12, 4.25 vs. 3.58, $p < 0.1$).

TILDA versus NONE: For the users that compared using Tilda versus using only Slack, 4/6 found Tilda to be better for keeping track of what happened during the discussion, and 5/6 found Tilda to be better for looking back over the discussion. One participant mentioned the hyperlinks in the summaries, saying:

> "*I loved how Tilda let you click on links to go back to the original messages instead of having to manually scroll through myself.*"

However, only half found the TILDA condition better for participating in the discussion, and 5/6 users found NONE easier to use. Users in the post-task surveys also rated Tilda as more mentally demanding than using Slack alone (3.83 vs. 2.83, $p < 0.05$). This is not surprising given that the TILDA condition explicitly involves doing more than the NONE condition. As to whether the benefits of Tilda outweigh the costs, 3/6 stated they would use Tilda again for the same task while 2/6 preferred just using Slack. One participant said:

> "*Slack...is easier to use just because there is less to keep track of, but for organization, Tilda is the way to go.*"

This suggests that the cognitive load introduced by Tilda might be worth it for more demanding tasks. Another user said:

> "*If I was working in a corporate or work environment and in project management, Tilda would be perfect.*"

	None	Doc	Tilda
Time Spent (min)	11:12 (5:32)	12:12 (8:12)	12:55 (6:25)
Grade Received (out of 7)	5.79 (1.07)	5.89 (1.05)	5.88 (1.03)
Experience (5=Very Good)	3.14 (1.03)	3.59 (1.15)	3.83 (1.01)
Felt Rushed (5=Very High)	2.57 (1.02)	**3.04 (1.19)**	**2.08 (1.21)**

Table 4.2: Results from Study 2, where new users familiarized themselves with conversations from Study 1 using the artifacts created, broken down by the three conditions. We report the average and σ for time taken on the overall task, grade that users received from completing comprehension questions, self-reported experience on a post-task survey, and self-reported feelings of being rushed on a post-task survey. Statistically significant differences are in bold.

Doc versus None: In comparison, the 6 participants in Doc and None conditions overall rated Google Docs more poorly, with only 1/6 users preferring the Doc condition for participating in the discussion, 3/6 for keeping track of what happened, and 3/6 for looking back over the discussion. Only one user preferred to use Google Docs and Slack again for the same task while 3/6 preferred to use just Slack. In discussions, users complained about fragmented attention in the Doc condition, with one person saying:

> "*If you have multiple tools open then it's not clear where all of the people, where their focus is directed to.*"

Users also disliked how information was scattered in both the Google Doc notes and the chat log, saying:

> "*If I come back to many ideas I don't remember where they came from. It causes mental distress.*"

Indeed, we observed some participants actually having some discussions in the Google Doc as they were editing it in real-time. We also noticed participants using copy-and-paste often to transfer messages from the chat log to Google Docs.

4.5.2 Study 2: Using Structured Summaries to Catch Up on Missed Conversations

In the second lab study, we conducted a between-subjects experiment to compare catching up on concluded conversations using Tilda summaries versus Google Docs notes or just the Slack chat log. To do this, we used the 12 artifacts created in the first study, including original chat logs as well as any accompanying Tilda summaries or Google Docs notes, and recruited new participants to look them over and answer comprehension questions about the discussions. We recruited 82 users (mean age 35, 28 female, 54 male) from Mechanical Turk, an online microtasking platform. Users were paid $3.25 per task and were required to have a 97% acceptance rate and 1,000 accepted tasks.

Experiment Design

There were 28 users for each of the three conditions of None, Doc, and Tilda, with half reviewing the Travel task from Study 1 and half the Story task. Before the study began, the first author used the task descriptions from Study 1 to create 7 comprehension questions for each task without looking at any artifacts, and then created a rubric for each of the 12 artifacts from Study 1. Users were given access to the corresponding Slack group and the Google Docs notes or Tilda summaries, which were located in a separate channel in the same group, if they existed. Users were not taught about Tilda except for an explanation that the hyperlinks in the Tilda summaries pointed to messages in the original chat log. At the same time, users were also given the 7 comprehension questions to answer in a survey form. There was no time limit for users nor instructions to spend a particular amount of time. After answering the questions, users filled out a separate survey about their experience, including NASA TLX questions about task load [132]. After the study, the first author graded each response out of 7 based on the rubric while blind to the condition. Two responses were discarded due to a score of 1/7 or lower, and three Tilda responses were discarded for self-reporting they were unaware of the hyperlinking feature despite the instructions.

Results

TILDA users felt less rushed than DOC users. We calculated how long users took by looking at time spent filling out the comprehension questions, with results in Table 4.2. While users overall spent the most time in TILDA and the least time in NONE, a one-way analysis of variance (ANOVA) test found that these differences were not significant (F=1.15, p=0.32), due to the high variation in time spent. From surveying users about their experience, users rated Tilda the highest, though these differences were not significant as well (F=2.41, p=0.09). Finally, we asked users about their task load, including the question of "*How hurried or rushed was the pace of the task?*" on a 5-pt Likert scale, where 5 is "very high". An ANOVA test yielded significant difference between the conditions (F=5.54, $p < 0.01$). Using a post hoc Tukey HSD test, we found that TILDA and DOC are significantly different at $p < 0.005$, with TILDA users feeling less rushed. In post-study comments, users described what they found hard, with one user saying in the DOC condition:

> *"I would have used Google Docs exclusively to answer the questions, but not all the information in Slack was there (and vice-versa)."*

Another user said in the NONE condition:

> *"The conversation seemed to be all over the place, there was no structure other than a group randomly chatting."*

People who used TILDA hyperlinks had lower load. Since we did not give a tutorial on Tilda, we were interested to see whether and how users in the TILDA condition would choose to use Tilda summaries. Only 4 of the users in the TILDA condition chose not to click the hyperlinks at all (NO-LINK), while 10 users used links often (HEAVY-LINK), according to self-reports. The remaining users said they used the links a few times. HEAVY-LINK users reported somewhat lower mental load (3 vs. 4.25, $p < 0.01$), feeling a great deal less rushed (1.3 vs. 4, $p < 0.001$), and feeling a great deal less irritated and stressed (1.6 vs. 4, $p < 0.005$). Heavy-Link users also rated their experience as better, and spent less time overall yet still received a higher grade than No-Link users, though these differences were not significant. However, it is possible that our findings could be due to self-selection bias as opposed to solely due to using links in the summary to dive into the chat log.

4.5.3 Field Study

We conducted a week-long field study with four teams that use Slack to have work-related discussions. This field study allowed us to observe how Tilda is used in practice by real organizations.

We recruited teams by posting to social media and asking colleagues to distribute our call for participation. For the study, we aimed to recruit a diverse set of teams that work in different areas. We also sought teams that communicate in different ways, including teams that are remote and predominantly rely on Slack as well as teams that physically sit together. Users were compensated $100 to participate in the study and have Tilda installed on their team Slack account for a week ($20 per day). We told teams that we would store and analyze metadata about chat conversations and Tilda markup over the time that Tilda was installed but no textual content related to the chat.

Team A is a 3-person academic research team that sits together but uses Slack to keep track of ongoing research projects. **Team B** is a 6-person software engineering start-up that is fully remote and conducts all communication via Slack. **Team C** is a 4-person software engineering team that is partially co-located and uses Slack to troubleshoot and share resources. **Team D** is a 3-person fully remote team behind an online news blog that uses Slack to coordinate writing and publishing.

Study Design

Before the study, for 3 out of 4 teams, the first author was invited into the team's Slack organization to install Tilda and instruct members on how to use Tilda. In the

	Active Users	Total Days Active	Channels with Tilda	All Chat Messages	No. Tilda Summaries	Total No. Tilda Items	Avg Tilda Items Per Summary	Avg Tilda Items Added Per User
A	3	6	6	277	15	53	4.5 (5.4)	20.3 (3.8)
B	6	9	4	870	40	220	5.5 (6.3)	35.7 (10.5)
C	4	5	6	478	22	101	5.8 (5.9)	31.3 (23.3)
D	3	8	9	373	36	51	1.5 (1.9)	17.7 (11.4)

Table 4.3: Overall usage statistics for the 4 Slack teams in the field study. Teams had variable usage of Slack as well as Tilda, with Team B as the most active overall.

case of Team D, the first author trained one individual in the team who then installed Tilda and taught the rest of the team on his own, due to the team's preference to keep their chat private. The training sessions were overall quicker than in Study 1, taking under 15 minutes using the same training materials, due to people's expertise with using Slack.

Participants were each asked to make a minimum of three notes or tags per day using Tilda, or 15 Tilda items in total over the course of the study. We chose to set the required activity low so we could see voluntary usage. We also gave no further requirements or suggestions so that users would be free to decide how to use Tilda. At the end of the week, 13 out of the 16 total number of users filled out a survey about their experience. At that point, we let the teams uninstall Tilda on their own, and three out of four teams continued to use it voluntarily for one to three more days during a second work week before we eventually took it down. We collected metadata on users' activity while Tilda was installed, including the kinds of Tilda items that users added.

Results

Teams were active in using Tilda to mark up their chat. We report overall statistics in Table 4.3. As can be seen, there was variable usage of Tilda across as well as within the teams, that generally corresponded to how active they were in Slack as a whole. Almost all users went over the minimum number of items on a daily basis,

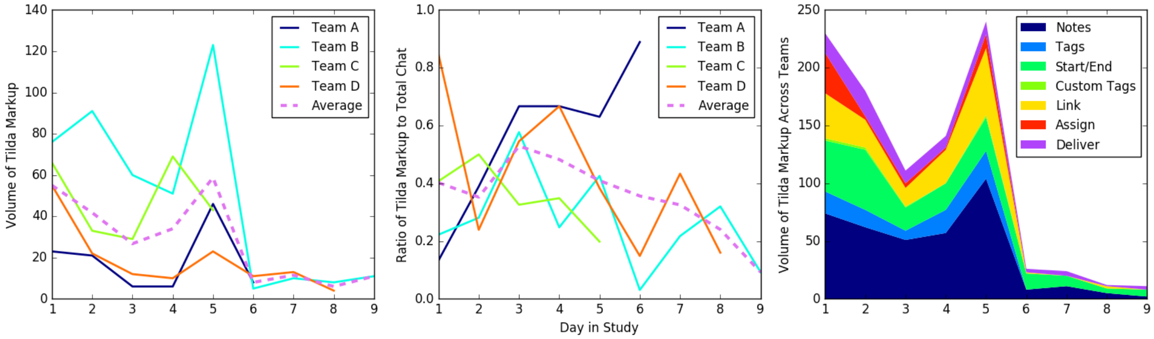


Figure 4-5: We show activity over the course of the study. On the left is the total raw volume of markup added to chat using Tilda by each team each day, showing high variability between teams and across days. In the middle, the volume of markup is normalized by the total number of chat messages sent by the team for each day. Overall, we see sustained activity for a few days before a gradual tailing off. On the right is the raw volume of markup across all teams per day broken down by markup type. Notable is the preference for notes over tags and significant use of linking.

and as mentioned, several teams used Tilda for longer than the required 5 days. The left side of Figure 4-5 shows the usage of Tilda over the course of the study, counting all possible markup that could be added to chat with Tilda. We remove days where there was no activity since some of the teams did not work on weekends. Different teams joined the study on different days, so the days of the week are not aligned. However, it was interesting that the peak activity was on different days for different teams. For team A and B, the peak was on the fifth day while it was the fourth day for Team C and first day for Team D. However, these fluctuations are perhaps a reflection of just overall activity in chat on those days. In the middle of Figure 4-5, we show the volume of Tilda markup normalized by the total number of chat messages posted in the team for each day. While these also fluctuate quite a bit, we can see that the average ratio for the four teams stays between 0.4 to 0.6 for around 6 days before decreasing. While we did not conduct a longitudinal study, the overall decrease in activity as the study concluded suggests that the tool will need to consider how to design for usage over longer periods of time. We present possible options further in the discussion.

As seen on the right side of Figure 4-5, notes overall saw higher usage than tags,

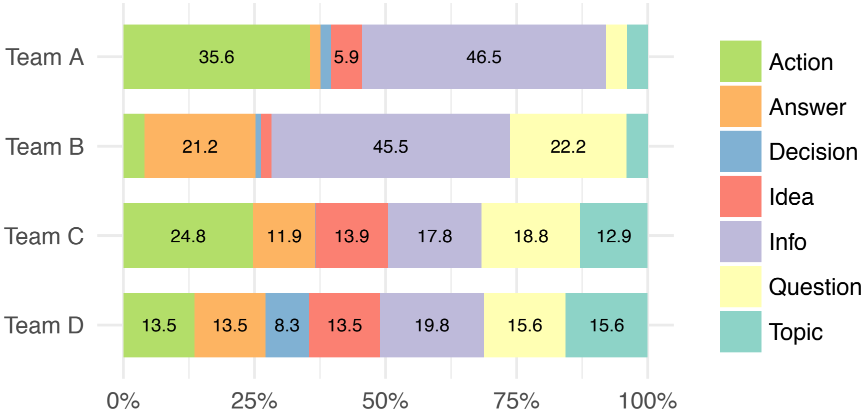


Figure 4-6: Percentage of Tilda items that were of each discourse act type for each team in the field study.

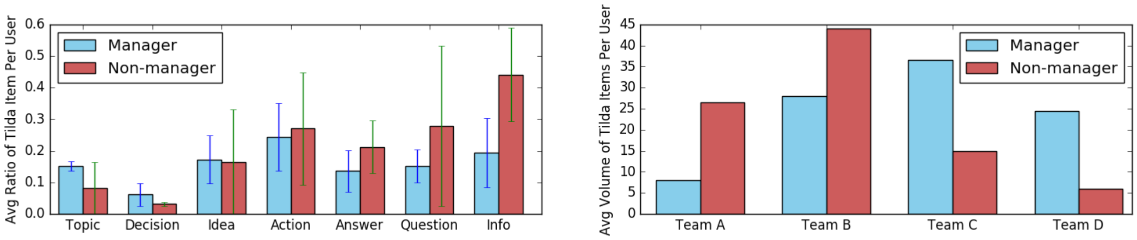


Figure 4-7: On the left, the ratio of Tilda items that were of each discourse act type for each user, averaged and grouped by managers and non-managers, from the field study. On the right, the volume of Tilda items added per user in each team, averaged and grouped by managers and non-managers.

thanks to Teams B and D, who favored notes almost exclusively, while Teams A and C were evenly split between the two. Across the board, custom tags were rarely used. One reason for this may be because the use of channels in Slack is already a decent separator of topics. Finally, there was surprisingly considerable usage of the linking feature as well as usage of the assignment feature earlier on in the study. From the post-study surveys, we asked users about their favorite feature, with several users mentioning the linking feature (3), Question and Answer items (3), Action items (2), and the automatic summary logging (3). In terms of missing or faulty features in Tilda, many users complained about how Tilda would take up too much real estate in the channel by posting (6), while 1 user wanted the ability to resolve Action items, 1 user wanted to link to multiple items, and 2 users wanted the ability to export the summaries to a document or their Trello board.

Teams and individuals personalized their use of Tilda to suit their needs. On average, except for Team D, users added around 5 items to each summary of a conversation, though this had high variance. Looking at the breakdown of items into their types, we can see in Figure 4-6 that Info was used frequently across all the groups, while Questions and Answers were used heavily by both software engineering teams. Other types were used more infrequently, especially Decision, possibly because users found the Action type more apt for the conclusion of conversations. We also asked users to rate each category for their usefulness in the survey; results were similar to Figure 4-6, with Action and Question rated the highest on average (4.3/5) and Decision and Topic rated the lowest (3.5/5). Overall, many users stated that they found the provided discourse acts expressive for all their notetaking needs. One participant said:

> "*The clear variety of different add actions was very useful; I didn't feel limited, like I had to shoehorn my types of choices into one box.*"

In Figure 4-7, we break down user behavior by managers versus non-managers, with management role self-reported from surveys. On the left-side figure, we break down the average ratio for different Tilda discourse act types for managers and non-managers. As can be seen, managers had a higher ratio of agenda-setting types such as Topic and Decision, while non-managers had a higher ratio for types more relevant to implementation details such as Info or Question and Answer. We also look at average volume of Tilda items left by managers versus non-managers, finding Team A and B had more non-manager participation, while C and D had less.

One interesting aspect of Tilda was how the norms of the team adjusted around the introduction of the tool. For instance, one participant described how adding Tilda improved the nature of conversation:

> "*I also noticed that Tilda changed our conversation flow (for the better). Since we were working with a finite set of tags...[the] messages...served a specific purpose that fell into one of the tags...questions were less likely to be lost...the action tag was a perfect way to remind us to take what we*

> *were chatting about and turn it into a tangible takeaway.*"

While we did not collect empirical data on this as we did not have access to prior activity in the teams, future work could analyze how the additional structure that Tilda allows might encourage certain types of discourse. Norms may also need to be set around the use of Tilda. One manager of Team B complained that the team left *too* many notes, leaving him to review unimportant information:

> "*I believe the summaries were useful and made it easier for me to review the notes as a manager. However, it hinged on the team being disciplined to only include important notes and this wasn't always the case...Overall I think if we got into the habit of using it effectively, it seems like Tilda would be big help to our workflow.*"

Perhaps some of the drop-off in proportional usage of Tilda by Team B starting from Day 3 was a result of decisions made, either implicitly or explicitly, to mark fewer items.

Tilda was effective for catching up and looking back. We were not able to capture data on reading of summaries due to limitations of the Slack API and so instead asked participants to self-report. Eight users said they used the summaries to catch up on missed conversations and rated the experience of catching up an average of 4.4 out of 5. One user said about catching up:

> "*Before Tilda I would try to scroll...This was very tedious...With Tilda this process was much smoother. I would usually check our Tilda responses channel and skim through the summaries to see what I missed. If a topic seemed interesting, I would expand it all and read through everything. If I was uninterested in the topic I would just move on.*"

Participants that were in the discussion also mentioned their motivation of marking up chat to keep absent team members up-to-date. One person who was on the partially remote Team C said:

> "*I work with a remote user a lot, and it was helpful to document what he needed to work on and clarify things he didn't understand.*"

Eleven people used the summaries to look back at old conversations they were in and rated their experience an average of 4.2 out of 5. Participants mentioned that a motivation for marking up chat was for themselves in order to keep track of things they needed to do or remember. One user remarked on using Tilda to look back through old conversations:

> "*Without Tilda - Scroll through or search for a keyword and try to find the message I think I remembered. If I can't remember or misremember something it can be frustrating trying to find it. With Tilda - Mark it and simply find the Summary either in the channel itself or the channel we had our responses in. Much less frustrating.*"

Ten users said they chose to set up a team-wide channel dedicated to summaries from the other channels, while 2 users chose to follow personalized summaries via their direct message with Tilda.

Tilda was used for structuring conversation and tracking important information. Some of the teams already had some mechanism for tracking longer term information and tasks, such as a Trello board or various Google Docs files. One person described how they liked having information tracked in one place, saying:

> "*Tilda gives us a somewhat better way to track information. It's useful to have everything all in one place...instead spread out like in Trello or Google Doc. Trello can get pretty messy easily...And I find our Google Drive directory hard to navigate...*"

However, some team members were used to the existing workflow they had with other systems and wished there was a way to sync them. Another participant thought of a separate site where summaries could be archived and searchable:

> *"...I really think that summaries should be exported/exportable to a different interface...for example to send to people off of Slack or to archive as a piece of important info...summary search...could be implemented on this page...For example, it would be nice if all action items could be pulled out*

> *to a running todo list organized by the topic of the conversation they came from (and linked of course).*"

For the teams that did not have mechanisms for keeping everyone on board and relied only on Slack, some members were excited about the additional structure that Tilda encouraged:

> "*We really didn't have a good system...Tilda made it muuuuch easier for us to fill someone in on something that happened...Overall I think Tilda greatly improved team communication over the week we used it. Conversations had better structure, team members were better kept up to date, and we actually had a way to save...results of our conversations for future use.*"

4.6 Discussion

Tilda markup adds structure to group chat conversations that can be beneficial to chat participants. First, in contrast to traditional notetaking tools like documents, Tilda's light-weight markup allows notetaking without forcing users to leave the conversation. As was suggested in Study 1, this approach offers a promising design pattern for making collaborative notetaking easier compared to alternatives. Study 1 also provides evidence that Tilda does introduce some mental load to users, but this could be a worthy trade-off for the organizational benefits it provides when it comes to discussing complicated things. Such benefits were echoed by participants in the field study who used chat extensively for work. In Study 1 and the field study, we also noticed high variability in how different users take notes, both in terms of note volume and their manner of notetaking. Similarly, we saw variability in Study 1 in how groups used the Google Doc to take notes, with different quality of outcomes. Tilda is more structured of a tool but still leaves room for variation, such as the number of items that make up a single summary. These observations suggest that, like good notetaking practices for documents, there may be some strategies to encourage better notetaking in Tilda. For instance, future iterations of Tilda could suggest closing a summary and starting a new one if many notes have been added, or asking users to pick the most important notes from a summary to create a higher level summary, in a recursive fashion [385].

When it comes to the output side, the field study echoed the results from the needfinding interviews, showing that catching up on or looking back over chat is a common task, and that it was improved with Tilda summaries. In Study 2 we found evidence that the links between conversations and notes were helpful for enabling newcomers to get up to speed more efficiently. As we observed in our field study, this structure was useful in providing additional context to conversations, allowing teams to organize and collaborate more successfully. Additionally, in Study 2, users felt less rushed using Tilda to catch up or look back over a separate document. This may be because the Tilda summaries are an alternative presentation or entry point for navigation into the original chat log and add no *new* content. In contrast, a document contributes new text and also leads to information spread out between two places. Given the use of Tilda summaries as a navigational tool, this suggests that an alternative presentation of Tilda summaries could have them overlaid or beside the original chat instead of posted to a separate channel or direct message.

4.6.1 Who Annotates and Why?

When it came to intrinsic motivation for users, we saw in the field study users mentioning that they added notes and tags in order to keep track of their own tasks and requests in the day, which then became helpful for other users. We also had examples of working with remote users in different time-zones where adding markup was helpful with asynchronous chatting. However, we did observe the importance of setting shared groups norms towards adoption of Tilda in our studies. A similar need for groups to get on the same page was expressed about group chat in general in our interviews, where some group chat users complained that inconsistent or non-reciprocal usage of certain features like threading sometimes led to even greater confusion. Even in the field study, we saw some people take many notes while others took only a few, though this could be because they did not use Slack or were not core members of the team. For Tilda to be successful, norms may need to be set by team leaders to motivate usage long-term.

In this work, our evaluations mainly focused on small groups of people conversing, and we did not explore how size of a team can alter the way Tilda is used. In a larger group, with hundreds or thousands of members, issues like social loafing, fear of participation, or contested norms [377] may be exacerbated. In such cases, an alternative design to Tilda's collaborative notetaking, reflected in earlier meeting bots like Debian MeetBot, could allow for the designation of an owner role for each meeting, who is in charge of adding notes, much like notetakers in live meetings. In some situations, such as in more ad hoc teams like Study 1 with no defined leader, this clear delineation of roles might be preferable. In future iterations of Tilda, the bot could also encourage participation by sending targeted proactive prompts to individuals to solicit notetaking.

Due to our decision to make Tilda a chatbot instead of an alternative chat system, we were constrained in the ways we could present summaries or messages to the group. This became an issue in the field study where the biggest complaint was about Tilda messages to the group taking up too much screen real estate. Due to these evident user experience issues, we chose not to pursue a longitudinal field study with the current implementation of Tilda. Additional deployments of Tilda could empirically examine alternative types of markup and summary presentations using short field studies or lab studies. In the future, a longer study on a new chat system where we have full control over presentation could allow us to further examine how norms and motivations around chat markup develop over time.

4.6.2 Towards Automatic Summarization of Chat

This work presents a first step towards a human-centered conceptualization of the goal of automatic chat summarization. In interviews, we collected empirical data around what kinds of summaries are desirable to chat readers, finding that structured summaries highlighting discourse acts were preferred over conventional presentations such as purely abstractive or extractive summaries. This result allows us to consider that the difficult problem of automatically summarizing chat conversation could potentially be tackled by breaking the problem down. Machine-learned models could augment the work that Tilda users do, such as by suggesting actions or simply performing some of them. Standard supervised machine learning techniques could be brought to bear on intermediate automatable problems include delineating separate conversations in a stream, labeling the discourse act of a message [382], finding messages that are candidates for tagging, linking messages to prior ones, and populating abstractive topic sentences or auto-tagging topics. These tasks have the benefit of reducing the learning curve and effort involved in using Tilda.

To build such models however, one must collect training data; luckily, Tilda too provides a path for fulfilling this role. More broadly, collecting rich training data can be a significant hurdle in developing models towards discussion summarization. In early pilots of our studies we conducted towards paid crowd annotation of public chat logs, we found that it was difficult for workers to make sense of a chat conversation they were not a part of. And as we saw in interviews, even if people are members of a group, it still takes effort to parse the back-and-forth when looking back over chat. Tilda manages this problem by making it possible to mark up chat conversations while taking part in them, when the conversational context is still fresh in their minds. In addition, we provide evidence that the Tilda system has value and direct benefits to users even in its current implementation as a primarily manual annotation tool.

4.6.3 Integration with Knowledge Management Tools and Email

Integration with outside knowledge management tools, such as wikis or documents, came up as feedback in both Study 1 and our field study. One could imagine Tilda chatting with existing bots or integrating with APIs to post to task management tools like Trello, Q&A sites like Stack Overflow, calendars [96], and code repositories. Likewise, one could imagine a website where additional organization of the summaries themselves could happen. Such an interface could be useful for newcomers looking to quickly make sense of the prior discussion in the team. Additionally, several interviewees described issues with triaging conversations that spill into both email and chat. Summaries could be inserted as embedded items in platforms such as email or forums that are more asynchronous. In all these cases, automatic links back to the original discussion in chat as well as automatic updating of content across links could manage the issue of information lost within multiple potential locations.

While Tilda bridges synchronous and asynchronous *access* of conversation, there are still questions about how to facilitate *partaking* in conversation for those who missed out. For instance, one person in our field study wanted a way to reopen a conversation that they had missed. This could be done by posting the summary to the relevant channel to remind users of the context and then writing a comment underneath. Any ensuing notes from the new conversation could get added to the original summary.

4.7 Future Work and Limitations

We have released Tilda as a public tool[2] and open-sourced the code[3], and aim to collect training data using Tilda towards automatic summarization tasks. Another area where we believe Tilda would be useful is for notetaking and summarization of video, audio, and in-person meetings, with the help of speech-to-text technology for transcription. Such a system could even work in concert with systems for crowdsourcing real-time transcriptions [196]. For instance, participants could collaboratively fix issues with transcription and highlight, tag, or vote on aspects of the discussion while conversing. While our work focuses on catching up and gaining an overview of a large chat log, we also uncovered issues that interviewees had with searching for particular items within chat. Future work could consider whether scrolling and other forms of orienteering behavior while searching [332] could be aided by signals left by Tilda. Currently, Tilda is a Slack-only tool; however because it was implemented using Microsoft's Bot Framework, it could be extended to other chat platforms that support

[2] `tildachat.com`

[3] `https://github.com/Microsoft/tilda`

bot integration with minimal additional development. The Slack features that we use, including emoji reactions, slash commands, and inline emojis, have uneven but growing support across other major chat platforms. For instance, emoji reactions are now supported in Facebook Messenger. Additionally, almost all platforms now support inline emojis, while slash commands could be simulated using hashtags.

4.8 Conclusion

In this work, we studied how users of group chat make sense of chat conversations when they need to catch up or look back, and we investigated how marking up chat messages to provide additional structure could help. From presenting 15 interviewees with different representations of chat information, we determined the importance of structure and discourse acts towards quickly understanding the gist and relevance of a chat conversation. From these findings, we developed Tilda, a tool for participants to mark up an ongoing chat conversation with various signals. The tool allows users to catch up on chat conversations through structured and linked summaries automatically created from users' notes and tags. From lab studies and field studies, we find that Tilda is effective for both taking notes and catching up on conversation.

Chapter 5

Murmur: Fine-Grained Moderation of Content Delivery in Mailing Lists

Mailing lists have existed since the early days of email and are still widely used today, even as more sophisticated online forums and social media websites proliferate. The simplicity of mailing lists can be seen as a reason for their endurance, a source of dissatisfaction, and an opportunity for improvement. Using a mixed-method approach, we study two community mailing lists in depth with interviews and surveys, and survey a broader spectrum of 28 lists. We report how members of the different communities use their mailing lists and their goals and desires for them. We explore why members prefer mailing lists to other group communication tools. But we also identify several tensions around mailing list usage that appear to contribute to dissatisfaction with them. We conclude with design implications that explore how to alleviate the tensions that we observe around mailing list usage and introduce a new system Murmur for fine-grained delivery specification within mailing list emails.

5.1 Introduction

Just four years after the invention of email, the first mailing list, MsgGroup, was created in 1971 to help Arpanet users discuss the idea of using Arpanet for discussion. In the 40 years since, mailing lists have become pervasive, helping communities share information, ask and answer questions, discuss issues, and build ties. More recently, alternative methods of group communication emerged, including discussion forums, Q&A sites, and social networking sites. As other tools gained prominence, some believed that mailing lists would die out and be replaced [191]. But mailing lists continue to be widely used.

Despite ongoing use, mailing lists have changed little from their original design. There have been some modifications and advancements, but generally mailing lists are used much as in the 1970s. While mailing list development stagnated, newer applications and websites have introduced numerous collaborative curation features, including following, tagging, and social moderation. These new systems and their features have been studied extensively in recent years. Email clients have also undergone dramatic changes in the last 40 years, so that now many people access their email in new ways [90].

Given the continued pervasive use of mailing lists, the lack of new development or research surrounding them, and advances in our modern social systems, we believe that a closer study of mailing lists today could reveal significant room for improvement. We consider the following questions:

- What are the reasons people continue to use mailing lists in the face of modern social media tools?
- What are the problems and limitations of mailing lists despite their continued use?
- How might we address these problems and limitations without ruining what makes mailing lists so attractive?

To gain insight into these questions, we studied the use of two mailing list communities through in-depth interviews. We augmented this qualitative examination with a survey of more members, and we additionally surveyed users from another 28 mailing lists of varying community types. We explored the diversity of goals, expectations, and perceptions among community members subscribed to lists, and how this can leave many users dissatisfied. In more detail,

- We saw significant disagreement over the preferred types, quantity, and tone of email delivered over each list;

- We found that many users muzzled themselves and others posted too much, based on their perception of others' preferences—perceptions that were often wrong;

- In particular, we found that the wide variation in how users handle incoming email influenced their perception of how the list should be used, to the detriment of others; and

- We observed that despite these problems, many users considered the mailing list superior for group communication to both web forums and social media.

Given these findings, we explore a design space for allowing diverse users to *all simultaneously* use the same mailing list in their different preferred ways without negatively impacting users with different preferences. Our results suggest that mailing list users could benefit from *greater flexibility and control* in how they choose their audience and their incoming content, and this might encourage more contributions that the community finds valuable. We also find a need for *greater transparency and social awareness* within mailing list systems to allow users to better know who their audience is and how their content is received. Our main contribution is an exploration of the current tensions existing within modern mailing list communities and opportunities so as to alleviate those tensions with design.

As a result of these findings, we develop Murmur, a reimagination of the mailing list that allows for fine-grained customization of content delivery by both senders and receivers. Murmur is implemented as mailing list software that can be used from any mail client as well as on the system's webpage. In this chapter, I will describe the features of Murmur and details of implementation. While Murmur has been publicly available and in active usage for several years, a formal study on the effects and usage of Murmur is slated for future work.

	Interview	**Survey**	**Subscribed***	**Posters****
Dorm	10	43	541	531
Lab	10	108	4,147	708

Table 5.1: The two mailing list populations studied in depth. *Number of subscribers at the time the survey was taken in May 2014. **Posters refers to the number of unique contributors in a period of 1 year starting from June 2013 to June 2014.

	Membership	**Archives**	**Posts/day***	**Moderated**
Dorm	restricted	public	15.75(13.53)	No
Lab	unrestricted	public	6.45(5.25)	No

Table 5.2: General information about the two mailing lists. *Average number of posts per day, followed by standard deviation, in a period of 1 year starting from June 2013 to June 2014.

5.2 Data Collection

We collected both interview and survey data, primarily relying on the qualitative interview data to gain a deeper understanding of the two communities we studied. The surveys, which reached a larger user population and a more diverse set of mailing lists, let us triangulate our interview findings.

5.2.1 Interview Study

We began in May 2014 with in-person interviews of members of two mailing list communities, summarized in Table 5.1 and Table 5.2 and characterized in more detail in the next section. The mailing lists were chosen because they were well established in terms of age and integration into their respective communities, giving them a sizeable membership, community participation ratio, and posting frequency that would allow for interesting dynamics to be observed. Our first mailing list, called Dorm, is for members of a 300-person undergraduate dormitory of a mid-sized U.S. university. We interviewed 10 (4 female, 6 male, median age of 22) members, including 8 undergraduates and 2 residential advisors. Our second mailing list, called Lab, is for members, affiliates, and followers of a 1000-person technology research lab in a different mid-sized U.S. university. We interviewed 10 (1 female, 8 male, 1 other, median age of 30) members, including 1 professor, 2 administrators, 2 researchers, 4 graduate students, and 1 former graduate student. Potential interviewees were recruited by emailing the target list, by emailing related mailing lists, and by word-of-mouth. We selected interviewees to reach a diverse set of users in terms of affiliation to the community, length of time in the community, and level of usage, including those who used the list infrequently or were unsubscribed.

Before the interview, we asked interviewees to reflect on their experiences and bring two posts or threads that were memorable in either a good or bad way in order to ground our discussion. We began the interviews by asking users about the posts they brought as well as their inclination or resistance to contributing in those instances. We then asked general open-ended questions about the mailing list, such as their opinions and participation level. We also had interviewees bring their laptops and demonstrate their strategies for organizing their mailing list email within their email client. Finally, we asked users to compare their mailing list with other community discussion systems that they used and to imagine what the list would be like if migrated to such alternative systems.

We employed a grounded theory approach [51] to allow themes to emerge from the interviews. They were conducted by the first author, lasted from 20 to 80 minutes, and were mostly open-ended to allow users to describe their experiences in detail. They were coded by the first author using standard qualitative coding techniques [232] to find concepts around what users liked about mailing lists, frustrating or rewarding experiences, and moments of doubt or self-censorship. The authors as a group iteratively discussed the codes and grouped them into themes. Some groupings were made from concepts that seemed contradictory; these form the tensions that we describe later. Others were made from commonly-expressed explanations for behaviors and preferences.

5.2.2 Survey

Using the themes generated, we then built a survey to see whether our interview findings could be confirmed by a larger subset of the two communities and by a more diverse set of mailing list communities. We built a 4 page web survey using SurveyMonkey that had a combination of multiple choice questions, free-response questions, and 5-point Likert scales. We deployed the survey after interviews concluded. In addition to Dorm and Lab, we surveyed 28 other mailing list communities. These communities were found by asking others to publicize the survey to mailing lists they used. We aimed to reach a diverse set of mailing list populations and selected communities of varying sizes and functions.

Our survey investigated users' attitudes towards and perceptions of their mailing list, which is why we relied on self-reported data. To build the survey, we took the themes developed and converted them to sets of questions, with some multiple choice responses taken from the codes extracted from the interviews. We asked users about their strategies for managing their mailing list email and characteristics of the list. We inquired whether they cared about things like missed email, irrelevant content, or high volume. We delved into how users felt about lengthy discussions and what gave them pause when considering posting. Finally, we asked users to rate potential changes to the list, including introducing hypothetical features to the list and moving to alternative systems.

We screened out 74 people who completed the survey in under 4 minutes, completed less than half, or had a variance below 0.5 for answers to Likert scale questions, which had items of reverse valence. Of 415 remaining participants, 43 (37% male, 56% female, median age 21) were from Dorm, 108 (67% male, 23% female, median age 27) from Lab, and 264 (33% male, 65% female, median age 21) from other lists. Some chose not to divulge their gender or age. The demographics for Dorm and Lab respondents reasonably approximate those of the membership. We did note a slight skew in gender towards more female respondents; however, we were careful to consider this in our analysis.

The total number of subscribers and unique contributors in the last year for Dorm and Lab are shown in Table 5.1. We presume that some email accounts were inactive or were filtered into a spam folder, but expect the number of people actually reading the mailing list is somewhere between the subscriber and unique contributor count.

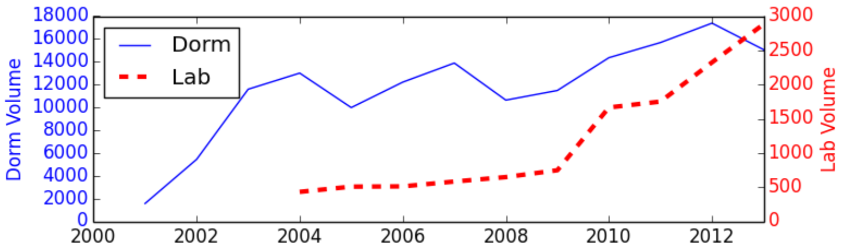


Figure 5-1: Total number of emails per year for Dorm and Lab.

Thus, we believe the real response rate to be above 5% for both communities. Our recruitment method of emailing the mailing list did not reach people who had left the list previously or did not check their email in time. This presents a non-response bias in our survey data, though we did take care to find and interview people who had left the list or did not check it frequently. We were able to reach these people by inquiring in person to members of both communities. We discuss potential biases in more detail in following sections.

5.3 The Mailing List Communities

We begin with a deeper look into what the communities that we interviewed and surveyed are like.

5.3.1 The Dorm Mailing List

As seen in Figure 5-1, the Dorm mailing list was started in the fall of 2001, with a general increase in volume in the years since. The community of Dorm is composed of primarily undergraduates and some residential advisors and staff that live together in residential housing. Students are randomly assigned to the housing community during their first years and stay until they graduate, so they generally know each other by name or face. Students are automatically added to the mailing list upon joining the community and are removed when they leave, though they can unsubscribe anytime. As Table 5.1 shows, the number of unique posters over the last year from June 2013 is quite close to the latest subscription number, meaning that almost all users posted to the list. However, about 25% of the unique posters only posted once. Interviewees described the content as comprised mostly of publicity for events organized by students for other students, with event announcements appearing several times a day. This activity is so prevalent that students name it “pubbing." This may account for why only about 30% of the posts in a year’s time were replies. At the time of our study, there were also many posts related to buying and selling items due to seniors about to graduate, highlighting the periodic nature of content due to the school year cycle.

5.3.2 The Lab Mailing List

The Lab mailing list was started in 2004 and has since seen a considerable increase in volume, also shown in Figure 5-1. The Lab community is composed of mostly current and alumni graduate students and some faculty, research staff, administrators, and undergraduates that are members of a technology research institute. Graduate students are automatically added but can unsubscribe anytime, and the list is public, so affiliates of the lab or interested parties may also be on the list. The volume is generally less than Dorm and varies less. As seen in Table 5.1, there are over 4,000 subscribers although only 708 unique people posted in a year’s time, suggesting that there are many lurkers and dormant accounts on the list. Of the people who did post, 51% only posted once. At the other end, the most frequent poster on the list posted over three times as much as the next most frequent poster. This person was referenced many times by name in both the interviews and surveys as a polarizing and outspoken list member. Interviewees described the list as a general-purpose list for the lab, with many job postings, housing listings, event announcements, and occasionally interesting discussions.

We additionally surveyed 28 other communities, with 3 sports teams, 9 extracurricular or cultural clubs, 5 academic groups, 6 dorms, 1 sorority, 3 social clubs, and 2 neighborhoods. Though we reached communities of different sizes and functions, many of them were connected to a university or were comprised mostly of university students due to our method of convenience sampling. We address the generalizability of our findings in light of our sample in a further section.

5.4 Why are Mailing Lists Still Relevant?

We first turn towards understanding why people still use mailing lists today in the face of modern discussion systems. Following this section, we will address problems related to mailing lists before discussing potential fixes to these problems, keeping in mind the positives we explore here. We report numbers in many cases primarily to describe our survey results but these numbers should be regarded as indicative due to our relatively low response rate. For all survey questions asking for degree of agreement with a statement, we use a 5 point Likert scale and code 1-2 as disagreeing with the statement, 3 as neutral, and 4-5 as agreeing.

We asked users to rate how often they used different group communication systems, including mailing lists, Facebook Groups, Google+ Communities, subreddits, or discussion forums. We found that after mailing lists, the next most popular tool for group communication was Facebook Groups. When asked about the Facebook Groups they were on, interviewees overall said that there was generally little activity and that they checked them much less frequently than their mailing list emails, even if they checked Facebook several times a day. Some interviewees mentioned that DORM and LAB in fact had Facebook Groups, but that they had low membership and were mostly dormant. When we refer back to Figure 5-1, we can see that volume has generally gone up over time on both mailing lists even during the growth of Facebook.

We asked users to imagine moving their mailing list to other systems and consider what would change. Overall, interviewees believed that moving their list to Facebook would result in less activity or discussion and preferred to continue using their mailing list. From the surveys, only a small minority of respondents liked the idea of moving their mailing list to a Facebook Group (13% agree, 15% neutral, 72% disagree). We found even lower percentages in favor of moving to a subreddit or a web forum (5% and 9% agree respectively). We now explore several differences we encountered in how people thought of email versus social media and how this played into their preference for mailing list communication.

5.4.1 Email Feels More Private than Social Media

A number of interviewees said that social media somehow felt more public than mailing lists, when explaining low activity on Facebook Groups. This was interesting because it was technically untrue; the mailing list archives were public in both lists while a private Facebook Group would not be visible outside the group. However, most interviewees of both groups, Dorm in particular, were surprised to find that the mailing list archives were publically accessible. We also found that both interviewees and those surveyed severely underestimated the number of people on the lists, echoing other research [20]. For instance, only 7% surveyed of Lab properly guessed how many people were subscribed to the list. Instead, the median guessed list size on the survey was 500-800 people, an order of magnitude lower. As another potential reason why, many interviewees commented that on Facebook people's identities were more tied to their messages because of the proximity of profile images and linked profiles:

> *There's a greater sense of [Facebook] being public...you can see everybody who's on there. It's very visible, very present. Whereas on email, you're sending it into the mystic...you don't see all the faces staring back at you.*
> –Dorm

Other interviewees reasoned that the archives were harder to access and read, while it would be easier to scroll down a group's Facebook page. This suggests that efforts to make archives more searchable or readable need also to clearly demonstrate to members the size of their potential audience.

5.4.2 Email is Still Considered Work, While Social Media is Play

To many interviewees, email was still considered more professional and more associated with information and productivity than social media. From the surveys, a majority of users thought that email was more professional than Facebook (61% agree, 25% neutral, 14% disagree). In a similar vein, interviewees associated discussion forums like Reddit with procrastination. When asked if the mailing list should move to Facebook, one user said:

> *I wouldn't be surprised if people start posting cat videos to this. [Facebook] has been a distraction for most people... When I look at* [Lab]*...I don't see it as a place to post cat videos.* –Lab

As we will explore later, a common theme we found across our communities was an enjoyment of and a desire for more interesting discussions within the mailing list. One interviewee who wanted more discussions felt that they would not thrive on Facebook, but they also did not quite fit with his perception of email as more work-related:

> *...it leaves the open discourse in an awkward split between personal conversation, Facebook Groups, and the part of email that's not all business-y.* –Dorm

While overall few users minded that group conversations were going into their email (14% agree, 22% neutral, 64% disagree), our data may be biased in that users who unsubscribed were less likely to respond to our survey. Some interviewees, one of whom had unsubscribed, indicated that the discussions in their inbox distracted them from typically work-related emails. Many interviewees similarly felt that some or more posts on the mailing list were not work-related. Given that these group conversations still felt more work-related than social media to many interviewees, this suggests that mailing lists could be designed to be something *in between social media and email.*

5.4.3 Email is More Used and More Controllable

Despite the popularity of Facebook, many interviewees mentioned knowing people in the community who were not on Facebook but used email. Several interviewees also stated that they checked their email more often than they checked Facebook, with a majority of the survey respondents agreeing (67% agree, 14% neutral, 19% disagree). Another difference between email and social media is that email, using the SMTP standard, is more readily customizable and viewable with many different interfaces. Many interviewees preferred to have the flexibility to set up custom filters, tags, or notifications. One interviewee expressed frustration with Facebook's interface, which is not customizable:

> *You only have the choice [on Facebook Groups]...I want to watch every message...or I don't. If you say yes, then...your cellphone [is] beeping every 5 minutes. If you say no, you're going to miss everything. There's no in between where once a day I can...see what's new.* –Lab

Additionally, all email is *sent* to all recipients, which may make it more likely for email to be *seen* as well, though this could depend on access methods as we shall explore later. In contrast, systems like Facebook that employ an opaque algorithm for displaying content make it difficult to even determine who receives what:

> *Facebook plays games with what they show people and so there's no even clear notion of who it is that's seen what you're sending...[Email's] really the only mechanism where it comes with this feeling of it'll get seen.* –Lab

In the survey, a majority said that they enjoyed having the flexibility and power to organize their email the way they wanted (67% agree, 22% neutral, 11% disagree).

5.5 Tensions Within Mailing List Communities

We now turn to examining several tensions that we observed within the mailing list communities we studied that may lead to problems. To facilitate our exploration, we categorize the types of mailing list posts into *transactional* (events, sales, etc.) and *interactional* (discussion, humor, etc.) communication. These categories have been used for spoken discourse and found in prior mailing list studies [131]. We

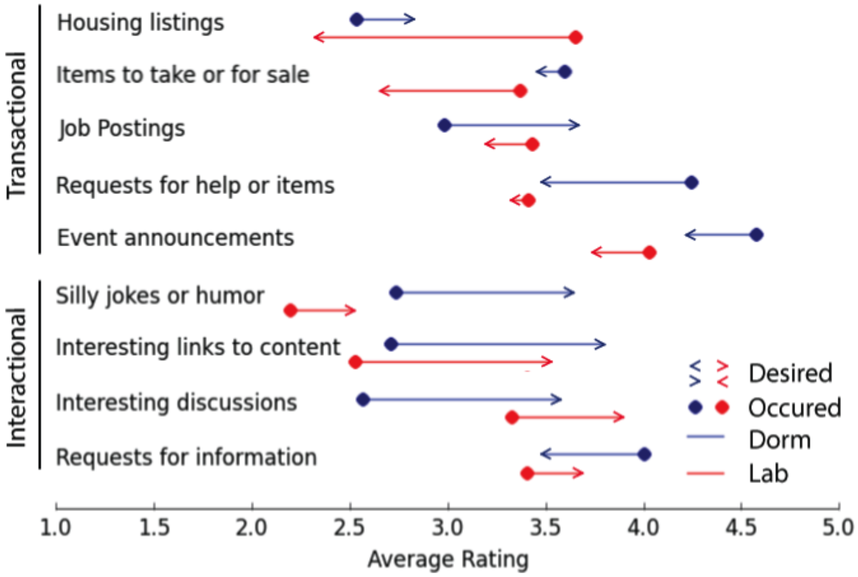


Figure 5-2: Survey results drawing an arrow starting from the number of times users stated a type of post occurred (circle) and ending at how often users would have liked for it to appear (arrow). Ratings are displayed as averages. Scale is 1–Never, 2–Once a month, 3–Once a week, 4–Once a day, 5–More than once a day.

acknowledge that not all posts fit easily into one category and that some intended transactional posts become interactional.[1]

Breaking down the mailing list content more finely, we asked survey respondents to self-report how often certain types of posts occur and also how often they would *like* certain types of posts to occur. In Figure 5-2, we visualize the difference for Dorm and Lab. To validate our survey results, two people were employed to manually tag 100 random emails from May 2014 from each mailing list into one of 9 categories we chose, given the subject line and body of the email (Cohen's kappa=0.70). After resolving disagreements through discussion, we found that with minor exceptions generally people's perceptions of how many emails they received of each category aligned with the normalized frequencies we found.

[1] For instance, our post to Lab soliciting participants for our survey turned into a multi-week, 70-post discussion on the ethics of using Amazon gift cards as a reward.

	Dorm	Lab
+	*"One of my favorite...types, is the sort of intellectual discourse...there was a golden time, where you had the right combination of people, you could get a good...intellectual discussion."*	*"I sometimes wish there were more meaningful conversations about technology and less about logistics and so on. ...Those things show up a lot in talks but I don't think people discuss them enough."*
−	*"I personally am glad that [the discussion is] gone. I think it keeps [Dorm] to be much more efficient."*	*"...if I had to pinpoint an ideal level for me, personally, I don't know, maybe 10 to 15 percent less of what [the discussion level] currently is right now, would be great."*

Table 5.3: Interview quotes expressing positive and negative feelings about interactional content on the two mailing lists.

5.5.1 Tension 1: Differences in Type of Content Desired

In our interviews, we found that often users even within the same community had different ideas about what their mailing lists should contain. For instance, we learned from the senior student interviewees of Dorm that the mailing list used to have more discussions during their sophomore year, because of a certain set of outspoken seniors. As seen in Table 5.3, interviewees disagreed on whether that was a good thing. Interviewees in Lab also disagreed on the optimum level of interactional content. When it came to more specific categories of email, such as the ones in Figure 5-2, interviewees also disagreed. Some interviewees were strongly in favor of more lighthearted humor or silliness on the mailing list while others were strictly against it. As another example, one interviewee spoke about job postings:

> *I think people will differ in that evaluation. I'm sure there's lots of people who actually appreciate the job postings and stuff whereas I'm not looking for a job.* –Lab

We asked interviewees about how to reconcile the amount of transactional posts they wanted on the list versus the amount of interactional content they wanted. Some users acknowledged the tension between the two functions of the list:

> *The users of* [Dorm] *are the types of people that don't really care about spammy stuff in their inbox ...Only upon reflection of what* [Dorm] *use to be, do I stop and think like yeah, maybe that spammy stuff kind of pushed out more of that intellectual conversation...* –Dorm

In the survey data, when asked if the mailing list should stick to informational posts, a sizeable minority agreed (24%). On the flip side, 34% wished the list would have more discussions. Because these two questions were inter-related (Cronbach's alpha = 0.70), we added them together to create an overall *discussion-desired* measure. The average variance within the communities was 1.95, demonstrating a large spread of preferences. As shown in Figure 5-3, we also saw that within Dorm and Lab, while there is a general trend towards wanting more discussion, there appears to be no one level of discussion ideal for even a large plurality of users.

Users also had very different ideas about how focused their list content should be. Some interviewees were sensitive to relevancy and stated that they would leave the list if there were an increase in the number of irrelevant emails:

> *...I don't like getting email. Especially when it's not applicable.* –Lab

Other interviewees didn't mind irrelevant content because they wanted to feel more connected to the community and liked knowing about things going on, even if it didn't affect them. Some also appreciated serendipity, or being able to stumble across information they normally wouldn't read:

> *There's always that case that there's an event or something that I'm like, "...This is really cool." I never would have found that, if it wasn't for* [Dorm]. *...To reduce those [types of posts] would be probably detrimental to those small instances.* –Dorm

> *To me it's kind of like a nostalgic, ambient awareness sort of thing...You still want to kind of keep tabs on what's going on even if you're not fully practicing everything.* –Lab

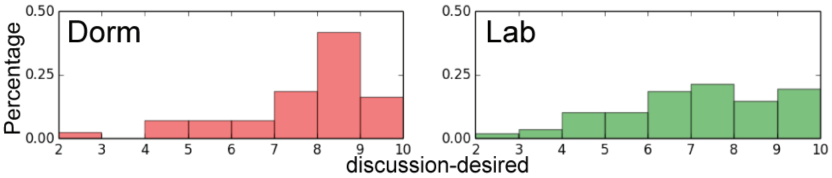


Figure 5-3: The *discussion-desired* measure, from 2–least desired to 10–most desired, for Dorm and Lab.

Interviewees also disagreed about whether replies to posts should appear on the list. The following two quotes are from different interviewees from Lab:

> *...You might as well just post it [to the list]. If they're not interested they can either skip over it, or quickly skim over it, or whatever.* –Lab

> *I despise it when people hit the reply to all button instead of the reply to button.* –Lab

As we saw, people even in the same community often have different ideas about what type of content should be on the list. This tension could cause backlash or the fear of backlash against certain posts or behaviors, leading to self-censorship, as we explore further in Tension 2. Thus we must be cautious when designing new features that impact the entire community unilaterally, which could hurt an important minority of members, and should instead strive to give users more control of what they get.

5.5.2 Tension 2: Desire for Interaction vs. Hesitation to Post

As can be seen from the arrows in Figure 5-2, the desired occurrence is higher than perceived occurrence for most of the interactional content, while the opposite is true for transactional content. Despite this general desire for more interaction, we found paradoxically that *many users who wanted more discussions did not contribute to them.* One interviewee acknowledged the discrepancy between her actions and desires, saying:

> *I think it is kind of a Catch 22...I want more discussion but I also don't want to put myself out there...* –Dorm

Post Type	Dorm		Lab	
Misinterpretation	3.79	(1.06)	2.69	(1.01)
Appearing Stupid	3.54	(1.19)	3.08	(1.14)
Heated Arguments	3.52	(1.14)	3.19	(1.19)
Privacy	3.50	(1.24)	3.10	(1.19)
Negative Voices	3.29	(1.24)	3.16	(1.22)
Strangers	3.26	(1.11)	2.74	(1.15)
Spamming Large Audience	3.05	(1.29)	3.41	(1.14)
Time and Effort	3.00	(1.11)	2.71	(1.15)
Presence of Authority	2.52	(1.12)	3.06	(1.26)

Table 5.4: Survey results for what gave users pause before posting for Dorm and Lab. Averages and standard deviations are reported. Scale is 1–Strongly disagree, 2–Disagree, 3–Neither disagree nor agree, 4–Agree, 5–Strongly Agree.

We found additional evidence of this in the surveys when we focused only on the respondents who said they wanted more discussion. Of the people who wanted more discussion, a majority of them had actually *never* participated in a discussion on the list (66%).

Though low levels of posting can be attributed to issues such as social loafing [165], we asked interviewees whether there were times when they wanted to participate in a conversation or had even written a post but did not send it. In these cases, users were *actively interested* in participating but were *deterred* for various reasons. We categorized the reasons that interviewees cited for why they self-censored their posts. Our survey then asked which of these categories gave respondents pause when posting. Though previous research has uncovered some similar deterrences [308, 345], we report them here for greater contextualization and as motivation for some of our later design implications. Below, we discuss a few of the categories that stood out in interviews and surveys. Full survey results can be found in Table 5.4.

Spamming Large Audiences: Many interviewees stated that they were worried about spamming a large number of people. This was the most troubling of all the issues for Lab and the other communities that we surveyed except Dorm. One interviewee talked about the times he wrote long replies but never sent them, saying:

> *I'm not sure what it is that I would be losing if I hit that send button but...I*

> *felt...I'm just spamming people...and I'm only perpetuating inbox overload to people.* –Lab

Misinterpretation: This issue was the most troubling of the issues for Dorm survey respondents. Many interviewees from Dorm said that they were hesitant to engage in discussions of controversial issues over email for fear of misspeaking and offending someone. This issue may have been more salient for Dorm because members live in close proximity and all generally know each other. To them it was difficult to craft a response that they felt would be politically correct and would not be misinterpreted:

> *Writing an email that is nuanced enough for* [Dorm] *without pissing people off just like takes so long that it's not worth my time.* –Dorm

Heated Arguments: People often expected that joining a discussion might lead to a heated argument they didn't want to get into. Interviewees from Dorm and Lab could both name particular people on the list that they felt were "trolls," or people who could be counted on to spark controversy, leading to repetitive arguments. This was the second and third most problematic issue for Lab and Dorm respectively.

> *...A couple times...I would feel like I had something to say and I would write this reply...I would spend a lot of time on it and then think this isn't worth it...it's just going to devolve into an age old argument of the same type that has happened over and over again.* –Lab

Appearing stupid: People from both communities were worried that they would be judged for appearing stupid. In Lab, this often translated to being afraid to talk about technical issues, for instance:

> *One thing that* [Lab] *is relatively devoid of is technical questions. You're keenly aware that the way...you'll ask a question signals an ignorance that you're afraid to show in* [Lab] *with such smart people.* –Lab

This was also the second most problematic issue for Dorm. One international interviewee said that she was embarrassed about her poor English and chose not to post questions, even when she really wanted help.

Summary: As can be seen, users were deterred from posting due to their fears of how their participation would be perceived by other members. We saw many users were afraid of spamming others with unwanted discussion, yet still a majority of people wanted to see more discussion. This highlights how these fears may sometimes be misfounded. Indeed many interviewees spoke of experiences where their participation resulted in positive outcomes. However, there were also times when users' fears were not misfounded, with arguments or harsh responses resulting. These conflicts may possibly be exacerbated in part to differences outlined in Tension 1.

5.5.3 Tension 3: Push vs. Pull Access

The last tension we explore is related to how users chose to access their email and how this may have affected their attitudes and actions on the mailing list.

Information access and exchange has often been differentiated as *push* versus *pull*. In push systems, senders actively "push" content to recipients, while in pull systems, senders make content available and recipients "pull" it at their leisure. There are two aspects of push systems that are often expected: recipients receive all messages, and they receive them in real-time. Neither of these expectations hold in pull systems, where recipients can ignore content or read it when they wish.

Traditionally, email is considered a push system while discussion forums and message feeds such as on Facebook, Twitter, or RSS are accessed more like pull systems. But while email *systems* are push-based, some users' email access *practices* have shifted to ways that are more typical of pull-based systems. We find evidence of this shift in users that use automatic filing to divert mailing list email away from their inboxes to separate folders. The difference in behavior of these users versus users within the same community that access their mailing list in a traditional way may lead to tensions. We also note that while users can digest their email, only 2% of those surveyed used it, so we do not focus on that group.

5.5.4 A Pull Experience via Automatic Filing

A currently popular practice is filtering email into secondary folders that are then accessed less often. Many interviewees said that they had their mailing list email automatically filed into a separate folder. One interviewee, explaining the difference between his interaction with his main inbox and his mailing list folder said:

> *It does change your interaction. It's a lot less urgent. I perceive* [Dorm] *as something that's less important. ...I want to check* [Dorm] *so I'm going to open it and look at it. ...Whereas with my email inbox it's coming there. I need to check this.* –Dorm

Another user had mailing list email come to his main inbox until Gmail began automatically filing it into a Forums tab:

> *Once Gmail made that change...most of my day is spent in the Important Email tab and I rarely look at the Forums tab. ...I think I skim* [Lab] *less not because of a disengagement from the list but just because the email client has suddenly hidden them...* –Lab

For some interviewees, filing the mailing list into a separate folder meant that occasionally they would forget to check it for an extended period. Others said that they purposefully checked the folder less often when they were busy.

> *It doesn't feel like I can't keep up perhaps because I don't want to be reading every single email.* –Dorm

> *I treat it the same way that I would treat a...water cooler where you walk by and there's some colleagues talking...but you can't spend all day at the water cooler.* –Lab

In our survey, automatic filing was the most popular strategy for Dorm and second most popular for all other communities (48% Dorm, 17% Lab, 11% others). This difference may be due to the relatively high volume of emails that Dorm receives. A majority of automatic filers reported that they did not mind missing email from the list (67% agree, 20% neutral, 13% disagree). A majority also reported that they read email from the list when they felt like it instead of when it arrived (70% agree, 15% neutral, 15% disagree). Additionally, many interviewees with this strategy stated that they did not mind irrelevant email or high volume from the mailing list specifically because the emails were being filed away.

5.5.5 A Push Experience without Automatic Filing

We observed a different attitude from the users who did not file their mailing list email separately from their normal email. These users had to go through each mailing list email just like any other email because it was arriving in the same place. One such user noted problems that arose when he neglected to read his mailing list email:

> *...[there's] the risk of missing important mails when I allow many to go unread because the state of being read or unread is less signaling...* –Lab

In our survey, receiving content in the main inbox was the most or second-most popular way of dealing with mailing list email (23% Dorm, 58% Lab, 68% others). In comparison to automatic filers, a smaller proportion of these users did not mind missing email from the list, with more than double the previous population disagreeing (50% agree, 31% neutral, 29% disagree). Fewer users also reported that they read email when they felt like it as opposed to when it arrived, with most disagreeing (36% agree, 16% neutral, 48% disagree). Interviewees also expressed a willingness to unsubscribe if email became more irrelevant or volume started to increase:

> *[Email] is something that I make an effort to stay on top of. I will unsubscribe from mailing lists if I think it's sending me too much email that's not relevant.* –Lab

Users who do not filter their mailing list may be more sensitive to irrelevant email, making it difficult to maintain a casual relationship to the community or benefit from serendipity.

5.5.6 Comparison

The survey data supported our association of automatic filing with pull-based and of manual handling with push-based behavior. We found that automatic filers were more likely to completely miss email from the list (18% vs 7%) and to not mind missing email from the list (67% vs 50%). Automatic filers also were more likely to only read mailing list email when they felt like it, not when the email messages arrived (70% vs 36%). On the other hand, users who had no filing strategy were more likely to read every message from the list (81% vs 75%). In the interviews with automatic filers, we found an expectation that others were accessing their email in the same way and would not mind additional email:

> *When...you get emails from someone, it doesn't take you that much time to just get rid of it. I think the people who really don't like spam already filter their* [DORM list]*... In which case my additional email really takes like three seconds of your time.* –DORM

An interviewee who did not have any strategy to differentiate his mailing list email had a very different thought-process when thinking of whether to post to the list:

> *When I'm thinking about sending an email to* [LAB]*, I'm like, "Wow, does every single person related to* [LAB] *really need to get this email?" If that's not the case, I probably wouldn't send it.* –LAB

Our interviews suggest that how people access their list email may impact how they feel about the list as a whole and how they then act as senders on the list. For instance, people who automatically file may assume that emails are not time-consuming and in turn may send more emails, annoying or overwhelming fellow members, or send less relevant emails, contributing to Tension 1. Conversely, people who read email from their main inbox may be more wary about spamming and thus not contribute as much, in relation to Tension 2.

5.6 Design Implications

Our results suggest that a better group communication system would keep the characteristics that people enjoy about mailing lists but also employ new features that alleviate the tensions we found. As simply moving to an alternative system outside of email may be unappealing, we instead consider incorporating features, some of which are common to social media or forum systems, within mailing lists. We note that several of our findings are not fundamentally tied to mailing lists but point to general preferences and tensions within communities, and thus they also suggest design implications for any group discussion system such as social media.

5.6.1 Coping With Diverse Preferences

As discussed in Tension 1, members of the same community often disagreed widely on what type and amount of content should be on the list. A potential way of fixing this could be to split the list into two or more lists, but this was rejected by many surveyed and interviewed, due to fears of it leading to less participation and splintering of the community. It is therefore essential to consider designs that permit many different sending and receiving preferences to coexist. One way to reconcile the tension over type is to allow tagging of emails with topics, as done by the Mail2Tag system for email [242] and systems like Reddit and StackOverflow.

Another popular feature on many social media websites today is voting on content. This feature allows crowds to collaboratively curate content; this can be used to promote the generally most upvoted content (social moderation) or to target content based on interests (collaborative filtering [185]). A sizeable subset of survey respondents were interested in receiving only the most interesting posts from their mailing list (33%). This feature can also be used to populate a point system for users to gain good standing in a community and as motivation to answer questions quickly, such as on community Q&A systems [340].

When considering such features, we must be cautious with changes if they destroy the "guaranteed delivery" implicit in emails' current push metaphor. This notion that email will be read, while no longer true as we saw in Tension 3, was still an appealing feature of email to many of our interviewees. We can maintain the status quo by pushing all content to a folder but still allow users to have certain messages pulled *to their main inbox* out of that folder, where they are more likely to view the content. As an alternative, we could have all users receive all posts as normal but alter subject lines to be prepended with interesting information such as tags or votes, which would allow users to skim them over more easily or conduct their own filing. For users that choose not to automatically file, we could digest, though we would need to carefully consider how to present time-sensitive emails and new replies to threads, as these were issues our digest users complained about. Finally, for the users who are unhappy with the level or type of emails and do not file or digest, we could simple not send certain emails. Clearly this would break emails' push metaphor and may lead to less participation from these users but may be a reasonable tradeoff for the users that we found that had already unsubscribed or were willing to completely unsubscribe if the volume or ratio of irrelevancy got too high.

Importantly, the new features we suggest would apply only to users who care enough to enable them and not to everyone. Also, in designing a new distribution mechanism, we must be careful to respect users' desire for control over what they receive. Algorithms for curating feeds may introduce biases and may be difficult for users to comprehend. Thus, schemes with *deterministic* filtering rules may be appealing for users who reject the opaque selection mechanisms provided by many of today's social media tools.

5.6.2 Overcoming Deterrences to Posting

Focusing now on senders of posts, we address the issues around users' deterrences from posting interactional content from Tension 2. In the surveys, a number of users stated that there were certain people on the list to whom they wished they could avoid posting (22%). By allowing users to exclude certain recipients from their post or block certain senders as is possible with some social media websites, this issue could potentially be alleviated. Blocked members may realize what is happening either through reading archives or by receiving a reply or forwarded email, and the poster would need to weigh that risk, just as with many social media systems. If the chances are perceived to be low, then the poster may find it worth it so as to not start a heated argument, have their conversation derailed, or otherwise avoid awkward or unwanted interactions. These features would need to be designed carefully so that members don't feel completely excluded, and the community maintains its cohesiveness.

Due to the lingering perception that all emails get read, many interviewees also had a fear of spamming large audiences and were deterred from posting. While deterrences can be useful to keeping volume down, this may also have the effect of silencing certain members over others as we saw in Tension 2 and 3, leading to a bias in participation. To combat this, we can destroy the notion of guaranteed delivery in only these instances to allow senders to send emails to a subset of the subscribers on the list. From the surveys, a sizeable portion of respondents said they would be more inclined to post if they could post to a random or targeted subset of the list (30%). We could allow the post to slowly propagate through the list of subscribers as it gains likes or replies, similarly to how content surfaces on some news feeds. While this makes receiving these emails no longer deterministic, this may be a reasonable tradeoff if the sender would have chosen to not post otherwise. Similarily, for users that are worried about looking stupid or being misinterpreted, we could allow users to send emails to a buddy or trusted group within the list to vet, moderate, or even edit the email before distributing it further. This could also potentially let the system mask the original sender without introducing the well-known problems around fully anonymous posting. This feature could be useful for other social media tools as well.

5.6.3 Reconcile Discrepancies Between Perceived Versus Actual

A common thread through our findings is a difference in people's perceptions of circumstances versus what they actually were. We discuss some potential ways to alleviate this with greater transparency. First, we found in Tension 2 that users were reluctant to post discussions, even though users generally wanted more discussion. To combat this perception, the upvote feature mentioned previously could also serve as feedback to let users know how their posts are received and encourage them to participate in the future. A significant portion of survey respondents thought favorably of adding upvotes or likes (34%). Another way to tackle discrepancies is to have users vote on the tags or common categories they find interesting and surface that at a community level. This could be juxtaposed with the frequency of tags going back a short period of time, much like in Figure 5-2, so that it is clearer what members want more or less of.

Tension 3 also uncovered differences in perception due to how others handle their email. To alleviate this, we could let users direct messages only to others who have not received too much email recently. As we are not able to find out the load of a user from their entire inbox, we would be limited to knowing the number of emails they receive from the various mailing lists they are subscribed to using this system. As entire organizations often use the same software for all internal mailing lists, this number may be a useful if not perfect measure of load for a member of such an organization.

Finally, this study brought up some privacy implications in that most people severely underestimated the number of subscribers to the list. Also, most did not know that the list was public to join or had completely public archives, yet many users preferred to use mailing lists over other social media due to a sense of privacy. These issues need to be made clearer to the users so that they are aware of who their audience may be. Additionally, any changes to make archives more readable or searchable or allow them to be crawled could have negative effects if not properly relayed to users. Even though some of these changes may cause people to self-censor even more, we believe coupling them with some of the other features we have mentioned may mitigate the effect.

5.7 Murmur: Reimagining the Mailing List with User Control Over Delivery

Given these findings, we explore a design space for allowing diverse users to *all simultaneously* use the same mailing list in their different preferred ways without negatively impacting users with different preferences. Our results suggest that mailing list users could benefit from *greater flexibility and control* in how they choose their audience and their incoming content, and this might encourage more contributions that the community finds valuable.

Drawing from our exploration into these tensions, we present **Murmur**[2], a mailing list system that aims to keep the benefits of email, such as greater confidence that messages will be seen, while introducing new features that are present in more modern systems such as Facebook, such as social moderation. Rather than using algorithmic curation, which puts the delivery of content in the hands of a model, Murmur allows users to have more explicit and fine-grained control to filter, block, follow, and otherwise curate how and whether discussions are received. Crucially, Murmur's features are not only targeted at receivers but also senders; for instance, senders may wish to limit their sending to a particular population, time period, or speed of propagation.

5.7.1 Murmur Design

Murmur is designed to give individual users greater ability to target the messages they send and the messages they receive in terms of person and topic, as well as define *how* they send and receive messages. There are also some administrator capabilities beyond that of a typical user but they are the same as common mailing list systems such as Mailman and not a focus of this system. Below are the possible actions that any individual user may take within Murmur. All settings are mailing list specific.

[2] murmur.csail.mit.edu

User Actions within a Murmur List

For each mailing list within Murmur, a member can set default delivery settings that apply to all messages within the mailing list. They can then make case-by-case decisions to override those default settings. Actions that have greater *specificity* take precedent over actions with lower specificity. As an example, a user may choose to block all emails, and then follows only the emails with a particular tag. From there, the user may be reading a particular thread with that tag and choose to block any potential replies. The blocking of the thread supersedes the following of the tag, which supersedes the blocking of all emails.

Default delivery settings: When a user joins or creates the mailing list, their default delivery setting as a receiver is set to *receive all messages.* They can change this setting so that by default, they *block all messages.* There is also the possibility to *receive the first post in every thread*, and not receive any replies unless they choose to follow the thread. Finally, there are further customizations along the dimension of time and grouping. Users can choose to receive all messages but in the form of a *digest*, similar to many existing mailing list software. They can also decide at what time intervals to receive messages.

Case-by-case receiver customizations: On top of the default delivery settings, users can then make case-by-case adjustments as they are writing and reading emails on top of those defaults at different levels of specificity. The type of adjustment they can make is predicated on the default settings that they selected.

First, users can choose to *follow/block another user's emails.* If they by default receive emails, they have the option to block; if by default they block emails, they have the option to follow. A user can also choose to *follow/block emails with a particular tag.* A customization at higher level of specificity is the ability to *follow/block a particular thread*, so that they receive the replies or they don't receive the replies, respectively. This customization supersedes others because of the greater specificity. Finally, receivers have the ability to tag a reply as well as upvote any email.

Case-by-case sender customizations: Beyond *receiver* delivery settings, users can also set *sender* delivery settings on a case-by-case basis. By default, senders are set to deliver their messages to every member of the list immediately. From there, users can *block sending to a user*, or mark specific people on the list that their emails should not go to by default. Users can also set their messages to be sent as *slow mail*. That is, instead of messages going to everyone at once, messages can propagate slowly through the network based on time or based on number of upvotes. Finally, senders can tag their email using a hashtag or using brackets in the subject line.

Ways to Perform Customizations

Murmur consists of two components: mailing list software and a web interface. Users can make changes to their customizations from within their email client of choice or through the web interface.

On the web interface, users can log on and join, leave, create, or deactivate a mailing list. They can also view information and logs of mailing lists that are public or where they are a member. For a user's given lists, members can post, upvote, and follow messages, as well as make other changes to their settings, such as review and edit the people, tags, and threads they follow or block.

Within a user's email client, whenever a message comes from a Murmur list, links are embedded into the footer of the email so that users can perform actions that are in relation to that message, such as upvote the message, block any further replies to the overall thread, or block future messages with the tag if the message is tagged. Users can also make changes to any of their settings as well as join, leave, create, or deactivate mailing lists by sending an empty email to Murmur and altering the email address for different actions. For example, to subscribe to a list, the email address would be `(group_name)+subscribe@murmur.csail.mit.edu`. Murmur then responds via email if the action was successful. There is a general purpose `help@murmur.csail.mit.edu` account to be reminded of the possible email addresses.

5.7.2 Murmur Implementation

The Murmur website is implemented using the Python Django framework connected to a MySQL database, though other databases supported by Django can also be used. For incoming and outgoing emails, Murmur uses Postfix, an open-source SMTP server, along with the Python library called Lamson, also an SMTP server, for integrating with the Django backend. Attachments are stored on an Amazon S3 bucket. Finally, features that must be run at regular intervals such as daily digest emails are configured using the Linux `crontab`.

5.8 Limitations and Future Work

Currently, while Murmur is available as a public tool and has been in use by several mailing lists for a number of years at this point, we have not conducted a formal field study of the tool. This is planned for future work with a large mailing list.

We examined only a student and a research-driven mailing list, and surveyed a convenience sample of other mailing list participants, many of whom were students or in academic roles. We did find the survey respondents for *Dorm* and *Lab* matched the demographics of the two primary communities we study but in general, we do not know our non-response bias. Generalization from our data should be done cautiously. Nonetheless, we believe we found some significant groups of mailing list participants who share the perceptions, expectations, and frustrations that we have outlined. In addition, because most of the participants in our interviews and surveys were young enough to have spent many of their formative years using social media, their preference for using mailing lists over social media for group discussion is potentially more interesting than that of a general population.

Many populations were not included in this study—for example, work-related mailing lists. There, different factors such as workplace hierarchy and the culture around socializing may play an important role. It would also be interesting to contrast mailing list usage with regular email and real-time work communications such as Yammer and HipChat. Similarly, we are curious how non-work systems might be enhanced or replaced by a real-time group chat interface, such as IRC. Several respondents also mentioned using GroupMe for small group discussion, which does for SMS what mailing lists do for email. It would be interesting to study how our findings translate to text messaging and whether GroupMe could then support larger groups.

5.9 Conclusion

Many people still use mailing lists to communicate within groups. Today, there are many new systems with new features for group communication, but they have not displaced mailing lists. We studied two mailing lists through interviews and surveys and surveyed 28 other mailing lists to understand how and why people use them and uncover important tensions within communities. We found that mailing list users within a single community *disagree on the types of content* the list should have; that despite wanting more discussion, *users self-censor* due to real and imagined concerns; and that *how users access* their mailing list email may alter their attitude towards receiving and posting messages. We also made a case for why simply moving to one of the new systems or building a new system outside of email may not be successful. From the issues we uncovered within current mailing list communities, we formulated design ideas introduced within a new mailing list system Murmur in order to alleviate the tensions we found.

Chapter 6

Squadbox: Friendsourced Moderation to Combat Email Harassment

Communication platforms have struggled to provide effective tools for people facing harassment online. We conducted interviews with 18 recipients of online harassment to understand their strategies for coping, finding that they often resorted to asking friends for help. Inspired by these findings, we explore the feasibility of *friendsourced moderation* as a technique for combating online harassment. We present Squadbox, a tool to help recipients of email harassment coordinate a "squad" of friend moderators to shield and support them during attacks. Friend moderators intercept email from strangers and can reject, organize, and redirect emails, as well as collaborate on filters. Squadbox is designed to let its users implement highly customized workflows, as we found in interviews that harassment and preferences for mitigating it vary widely. We evaluated Squadbox on five pairs of friends in a field study, finding that participants could comfortably navigate around privacy and personalization concerns.

6.1 Introduction

The internet has made remote communication frictionless, allowing people to interact from afar with strangers on a variety of platforms. While these powerful capabilities have in many ways been positive, they have also empowered bullies and harassers to target others like never before. According to recent reports by Data & Society [203] and the Pew Research Center [77], nearly half of internet users in the United States have experienced some form of online harassment or abuse.

Unfortunately, solutions for combating online harassment have not kept pace. Common technical solutions such as user blocking and word-based filters are blunt tools that cannot cover many forms of harassment, are labor-intensive for people suffering large-scale attacks, and can be circumvented by determined harassers. Even so, platforms have been criticized for their slow implementation of said features [130, 352]. Recently, researchers have built machine learning models to detect harassment [37, 167, 369], but caution that such models should be used in tandem with human moderators [6], due to the possibility of deception [147] and presence of bias in training data [22]. Indeed, paid human moderators already make up many of the reporting pipelines for platforms [227], but they still often fail to understand the nuances of people's experiences [26] and make opaque or inconsistent decisions [261, 353].

To devise better solutions, we examined the emergent practices of harassment recipients and systems designs that would better support their existing strategies. From a series of interviews with 18 people who have experienced online harassment, we learned about the nature of their harassment as well as how they cope. Interviewees came from a wide array of roles, from activist to journalist to scientist, and have faced harassment on a variety of platforms. Without existing effective solutions, we found that harassment recipients often turn for help to friends, who they can trust to understand their desires and maintain their privacy, using techniques such as giving friends password access to rid their inboxes of harassment or forwarding unopened messages to friends to moderate.

In light of these existing practices, we consider how to design tools that more effectively facilitate *friendsourced moderation* as a technique for combating harassment, a challenge that requires understanding differing individual requirements and managing potentially sensitive data. We present Squadbox, a tool that allows users to coordinate a "squad" of trusted individuals to moderate messages when they are under attack. Using our tool, the "owner" of the squad can automatically forward potentially harassing incoming content to Squadbox's moderation pipeline. When a message arrives for moderation, a moderator makes an assessment, adding annotations and rationale as needed. The message is then handled in a manner according to the owner's preference, such as having it delivered with a label, filed away, or discarded.

In the design of Squadbox, we embraced a philosophy that one of our first interviewees suggested and that later interviewees reaffirmed: "*Everything should be an option*". Perhaps the most significant takeaway from the interviews was that, as cases of online harassment vary greatly, no one particular solution will work for everyone. Some wanted to have access to all or some harassing messages; others did not. Some wanted their moderators to have greater power, while others wanted lesser. Some wanted to engage with harassers, and some did not. Thus, rather than making decisions for users about how exactly to use the system, we designed Squadbox to be highly customizable to different possible owner-moderator relationships and usage patterns. At the same time, we aim to *scaffold* the owner and moderator actions so they can be performed more easily than current jerry-rigged approaches. Our initial implementation targets email, as this is a platform that is particularly weak on anti-harassment tools but also one whose standard API makes it very easy to manipulate. The system can be extended to any communication platform with a suitable API, and we plan to do so.

We demoed the tool to five harassment recipients, receiving positive feedback on its current direction, in preparation for a public launch. We also conducted a field study with five pairs of friends that use Squadbox for four days, in order to study technology-mediated friendsourced moderation in a natural setting. We found that the use of friends as moderators simplifies issues around privacy and personalization of users' workflows. However, it also raised other issues related to friendship maintenance, such as the need to ensure moderators feel adequately supported in their role by owners.

6.2 Related Work

6.2.1 Online Harassment Research

There has been a great deal of work characterizing online harassment as a significant problem affecting many internet users [77, 203], with certain groups such as young adults [358, 364], women [95, 298, 358, 346], and those who identify as LGBTQ [203] bearing a greater burden. Research has found that 17% of internet users have experienced denial of access through means such as receiving an overwhelming volume of unwanted messages, having their accounts reported, or Denial of Service (DoS) attacks. Of all recipients of harassment on the internet, 43% have changed their email address, phone number, or created a new social media profile due to harassment [203]. As a result of harassment, many recipients simply withdraw from public online spaces [95, 346] or self-censor their content online [203]. Researchers and internet activists have studied or called for better processes to deal with harassment on various platforms [130, 227, 261]. Other researchers examine government policy on online harassment, finding it ineffective [222]. Researchers have also suggested design interventions for platforms to undertake, resulting from content analysis [293], interviews and surveys [346], and design sessions [14] with harassment recipients.

6.2.2 Community-Based Systems for Combating Harassment

By building on prior research methods and findings [84, 226], socio-technical systems researchers can play a part in mitigating online harassment through the development of novel systems. However, many researchers do not have access to the inner workings of platforms, which is often necessary to build or study possible interventions. Despite these limitations, we can look for inspiration from grassroots efforts by volunteers who have developed community-based anti-harassment tools [102]. Some of these tools include BlockTogether [143] and Good Game Auto Blocker [129], where users collaborate on shared blocklists of harassing Twitter accounts. Other community-based efforts include projects such as Hollaback! that elevate victims' stories [69], and systems such as HeartMob that provide a network of volunteers to support, provide validation for, and take action on behalf of harassment recipients [26]. The success of these tools suggests that a fruitful path forward for system builders may be towards empowering individuals facing harassment to better activate their existing communities. We take inspiration from this prior work in our approach to designing and developing Squadbox. We also take inspiration from participatory design processes [14] by learning from harassment recipients' existing strategies to then design a tool to augment those strategies.

6.3 Experiences, Preferences, and Strategies for Combating Harassment

We begin by investigating the nature of people's experiences with online harassment, their existing strategies for combating it, and how their personal support networks can play a role.

Through social media, professional networks, and cold-emailing people in the news, we sought out people who had experienced online harassment on *any* communication platform. 18 interviewees participated in a 45-minute to one hour-long interview with the authors via video, phone, or in-person. 12 had experienced harassment through email. The first half of each interview focused on understanding subjects' experience with harassment: the who, where, and how, as well as the impacts the harassment had on their life and actions they had taken in response. In the second half, we turned to discussing if and how subjects would use a friendsourced moderation tool. The first two authors performed a qualitative analysis of the interview transcripts, using a grounded theory approach to code the data and develop themes. In order to protect the identities of our subjects, some details and quotes have been edited, and we use "they" and "their" as personal pronouns for all subjects. Sixteen of 18 participants completed a survey to gather demographic information. Respondents ranged in age from 18 to 52, with an average of 33.25. Eleven identified as female, two as male, and the remaining three as genderqueer, non-binary, and a non-binary trans woman. Twelve identified as white, three as Asian, and two as Middle Eastern or North African. We group subjects and label their quotes using high-level categories based on the nature and sources of their harassment (elaborated in Table 6.3).

Occupation (Label)	**Platform(s) Harassed**	**Nature of Harassment**	**Peak Vol. per day**	**Avg. Vol.**
Graduate student (Res1)	Facebook, Twitter	Harassed via Twitter and private FB messages for sharing opinions on social issues, politics in academic circles.	10+	~1/month
Professor (Res2)	Email	Severely harassed for short period for controversial research.	50+	<1/month
Professor (Res3	Twitter	Harassed by an individual due to a fallout over a collaboration.	10+	<1/month
Scientist (Ex1)	Email	Harassed by an ex-significant other. Can't block, need to coordinate to avoid one another and not violate restraining order.	1+	~1/month
Director (Ex2)	Email	Was harassed and threatened by former significant others.	50+	~1/month
Librarian (Ex3)	Email, SMS	Harassed by an ex-significant other over the course of many years. Can't block, need to coordinate care of children.	10+	~1/day
Game developer (Fan1)	Email, Twitter	Harassed over several months by an individual pretending to be a fan. Also receives personal attacks on Twitter.	1+	<1/month on email, 50+/day on Twitter
Activist (Act1)	Email, Facebook, Twitter	Harassed on Twitter and FB because of activism on controversial and identity-related topics, and on email by ex-coworker.	50+	1+/day on email, 50+/day on Twitter
Activist (Act2)	Email, Facebook, Twitter	Harassed on Twitter because of writing and political activism.	50+	~1+/day on Twitter
YouTube personality (You1)	Email, Twitter, YouTube	Identity-based attacks and threats based on the content of videos. Has been doxed.	10+	50+/day on YouTube and Twitter, ~1/day on email

YouTube personality (You2)	Twitter, YouTube	Identity-based attacks and threats based on the content of videos. Has been doxed.	50+	50+/day on YouTube and Twitter
YouTube personality (You3)	Email, Twitter, YouTube	Identity-based attacks and threats based on the content of videos. Has been doxed.	50+	10+/day on YouTube and Twitter
YouTube personality (You4)	Facebook, Instagram, Twitter, YouTube	Identity-based attacks and threats based on the content of videos. Has been doxed.	50+	10+/day
Journalist (Jour1)	Email, Twitter, SMS	Harassed because of investigations conducted. Included fake website taunting and threatening the subject.	1+	∼1/month
Journalist (Jour2)	Email, Twitter	Harassed by people with dissenting opinions for political opinions in newspaper columns. Personal attacks and insults, some threats.	1+	∼1/day
Journalist (Jour3)	Facebook, Instagram, Twitter, YouTube	Large volume of harassment for a short period after being mistaken for someone controversial. Personal attacks.	50+	∼1/day
No response) (Spoof1)	SMS	SMS spoofing - both received messages, and messages sent pretending to be this person. Unclear who is the harasser.	1+	(No response)
Public Figure (Pub1)	Email, Twitter	Large volume of continual harassment, including greater waves due to public appearances. Personal attacks and death threats.	50+	(No response)

Table 6.3: Interview participants, labeled and grouped based on the nature and trigger of their harassment into groups around research (Res), ex-significant others (Ex), fans (Fan), activism (Act), YouTube videos (You), journalism (Jour), SMS spoofing (Spoof), and being a public figure (Pub).

6.3.1 Understanding Harassment and Mitigation Strategies

We first describe the nature of our subjects' harassment, how subjects communicate in the face of harassment online, and strategies that they have devised to combat harassment.

Harassment Defined by Content, Volume, and Repetition

Individual definitions and experiences varied greatly [309]. But in terms of message content, subjects described harassment as a personal attack, sometimes about aspects of their identity. They found these messages to be emotionally upsetting and draining. However, even when messages were not harassing at face value, they could become harassing when sent in high volumes, or when individuals made repeated, persistent attempts at contact despite being ignored or asked to stop. One interviewee, highlighting the oftentimes persistent nature of harassers, said:

> "*If I ignore their message, they'll send one every week thinking I'm eventually going to reply, or they will reply to every single one of my tweets.*" [You4]

Encountering Harassing Content Disrupts One's Day-To-Day

Subjects described being disturbed during their day-to-day activities by upsetting content, and expressed frustration at their lack of agency to decide whether or when to confront harassing messages. One subject said:

> "*Getting a [harassing] email when I'm looking for a message from my boss—it's such a violation. It's hard to prevent it from reaching me. Even if I wanted to avoid it I can't. I can't cut myself off from the internet—I have to do my job.*" [Act1]

Ex3 described how their harasser purposefully sent more harassing emails when they knew Ex3 was at an important event. Others talked about notifications, saying:

> “*The constant negativity really got to me…having it in your mind every 30 minutes or whenever there’s a new message…It just wears me down.*” [You4]

Volume and Nature of Harassment Impedes Communication

Even with a low volume of harassment, interviewees still found it affected their communication. For instance, Spoof1’s communication channels broke down completely when they became unable to distinguish between legitimate messages from friends and spoofed messages. For other interviewees, it was simply the massive volume of harassment that impeded their communication, echoing prior work on Denial of Service (DoS) attacks [203]. Sometimes, this harassment was incited by someone with a large following, who could direct “hate mobs” at will. As a result, harassment was often bursty—for example following publication of a controversial article—and thus many subjects alternated between spikes of heavy harassment volume and periods with little or no harassment. When subjects were inundated, many were left unable to respond to legitimate communication, such as from fans, their community, or professional contacts:

> “*It’s made it harder to find the people who genuinely care, because it’s hard for me to motivate myself to look through comments or…go through my emails. Why should I look through hundreds of harassing comments to find a few good ones?*” [You3]

The attack on their communication channels meant that some missed out on opportunities as a result of harassment. For instance, Jour3 mentioned missing an interview request amidst a flood of harassing tweets.

Platform Tools of Block, Filter, and Report are Inadequate

Nearly every subject we interviewed stated that they had blocked accounts on social media or email, though most felt this was not very effective due to the number of harassers and harassers’ ability to circumvent blocking. One said:

> "*Every time he makes a new email, he creates a new name as well...Not only new names, but he also pretended to be different people.*" [Fan1]

Others needed to see messages from their harassers, such as for coordinating childcare with an ex-partner (Ex3) or to be aware of incoming threats. Another reason subjects wanted to see messages was to get an overview of dissenting opinions, even their harassers', for work purposes (Jour1, Jour2). Finally, some subjects wanted the ability to track their harassment over time in response to their public activity (Pub1) or do damage control after defamation (Res3). Word- or phrase-based filters were also inadequate. Some subjects expressed frustration at the difficulty of coming up with the right words to block or managing changes in language over time. One described filtering out messages despite false positives, saying:

> "*I have suicide as a filtered word because I get more comments from people telling me to commit suicide than I get from people talking about suicide...If I have the energy to, I'll go through my 'held for review' folder to look through those.*" [You3]

Finally, nearly every subject had reported harassers to platforms and strongly expressed dissatisfaction with the process and the platforms' opaque responses. A common frustration was that the burden of filing a report was too heavy, especially when there were many harassers. Beyond platform tools, subjects also tried seeking help from law enforcement; the prevailing sentiment was that this was a time-consuming, fruitless experience, echoing prior work [222].

Harassment Works to Silence and Isolate Recipients

Subjects described self-censoring as a way to give harassers less ammunition with which to harass them, echoing prior work [203]. Res1 described blaming themself when something they posted led to harassing messages:

> "*It started changing some of the things that I would post. Now, [when] it happens I view that as, oh, I posted something I should've deleted.*" [Res1]

Another strategy subjects undertook was to make themselves harder to contact by closing Twitter direct messages from people they do not follow, not giving out their email, turning off notifications, or disabling comments. While this helped to mitigate harassment, it also made it more difficult to engage with people they did want to talk to—people they already know as well as non-harassing strangers, like collaborators, fans, clients, or sources:

> "*It's impossible to contact me if you don't have my contact info...I can't be available to journalists as a source...I used to get all these awesome opportunities and I just can't get them anymore.*" [Act1]

Asking Friends for Help can Mitigate Harassment Effects

A majority of subjects mentioned reaching out to friends or family for support and assistance. Act1 said that their best friend had their Twitter and Facebook passwords, and would log into their accounts and clear out harassing messages and notifications and block users. Ex1 said their spouse would log in to their email account and delete harassing messages, and Res2 had others in their department going through their emails. You4 said that their significant other would go through the comments on their posts and read aloud the positive and encouraging ones. Multiple subjects such as Act1 and Ex2 said that they would forward potentially harassing emails unopened to friends for them to check and forward back.

Summary

From analyzing our interviews, we determine several user needs that current platforms do not address. Users need to be able to divert harassing messages from their inbox or platform equivalent (**N1**), they need to be able to maintain private and public communication in the face of harassment (**N2**), they may need to ramp up or down mitigation strategies as harassment comes in waves (**N3**), they at times need to be able to read or get an overview of their harassing messages (**N4**), they need help managing blocklists and filters over time (**N5**), and they need help collecting and documenting harassment for official reports (**N6**). Meanwhile, the most effective strategy interviewees mentioned is asking friends for help.

6.3.2 Understanding Preferences for Friend Moderation

We saw from interviews that many already make use of a *friendsourcing* strategy to moderate their messages, albeit in an unsystematic way. Thus, we also spoke to subjects about actions friend moderators could take to help them and how tools could enhance their existing friendsourcing strategy.

Potential Friend Moderator Actions

Tagging and summarizing messages: One finding was that sometimes subjects wanted to read or learn more about their harassment (N4), though people had different preferred strategies. Some wanted moderators to tag their harassing messages so that they could divert them to a folder, and decide on their own when to open them (N1) or track categories or specific people over time for reports (N6). Subjects wanted tags about information such as subject matter, severity, and type of harassment. Similarly, they felt it was important that messages that might need escalation or a response be marked separately as urgent and sent immediately to them. Subjects had different ideas about what needed escalation, from "doxing" (publishing their home address), to death threats, to the harasser revealing other personal information about the subject. Others wanted a moderator rationale, summary, or redacted version of the message, so they could glean main points from the message without having to read the original harassing message.

Collaborating on word- or sender-based filters: Multiple subjects felt it would be helpful for moderators to collaborate on word-based filters that would flag a message for moderation or for automatic rejection (N5). Remarking on the cat-and-mouse nature of keeping filters up-to-date, one subject said "*People...know there'll be a blocklist, and they know...that they have to start spelling things funny or doing all this stuff to get outside of the filters...it needs to constantly be morphing*" [You2].

Similarly, subjects were interested in having moderators help build their sender-based whitelists or blacklists, similar to shared Twitter blocklists. Some felt that moderators should manage the lists, while others wanted a process where moderators could only suggest edits to the lists.

Responding to harassers: Subjects had mixed opinions about having moderators communicate with harassers. Some thought that being told to stop by someone other than the recipient could be impactful, or that moderators could diffuse the situation. Other subjects thought that moderators could help educate harassers. On the other hand, some felt communicating with harassers might be unproductive and actually lead to further harassment, citing the common refrain: "Don't feed the trolls". Overall, people had different ideas about if and how they wanted to view their harassment, how much power moderators should have to edit filters, and whether moderators should respond to harassers.

Privacy Concerns With Friend Moderators

Recipient Privacy: Subjects generally preferred friends as opposed to paid or volunteer strangers as moderators. This was due to privacy concerns regarding personal messages, as well as the inability of non-friends to understand their unique situation and preferences. One subject said:

> "*I feel like getting harassed is such an emotionally fraught experience that I prefer to turn to friends for support...it almost feels more violating to have somebody who doesn't know me read those...I would worry about personal information.*" [Act1]

Most subjects could name friends or family members whom they could trust to perform moderation duties or that had already helped them this way. Even so, most subjects were still able to name types of messages that they would prefer even friends not see—for example, those containing sensitive financial information.

Sender Privacy: Additionally, there are privacy considerations from the perspective of the sender, who may be unaware there is a moderator, even though the recipient is always capable of screenshotting or forwarding their message. One mitigation strategy would be an automatic reply to any initial message, notifying the sender about moderation and giving them a chance to revise or rescind their message. Some subjects felt this level of transparency could preserve privacy or even discourage harassers. Others preferred to obfuscate their use of moderation, as it might attract attention, leading them to be harassed more on another platform or make their harassers more determined:

> "*The second that someone knows that you're blocking people on Twitter, everyone tries to get blocked. As soon as someone knows that you're filtering out their emails, everyone wants to try to break your filter.*" [You4]

Moderator Burden and Motivation

Subjects were concerned about the workload for moderators. One stated:

> "*I feel guilty asking for too much help, which I think is just a problem a lot of people have when they're going through this.*" [Act1]

Subjects suggested features to alleviate this such as an on-off switch for the moderation tool, a rotating team of moderators, or the ability for moderators to set limits on their moderation. Others suggested a reciprocal relationship where they could moderate their moderator's emails, or join a group where everyone moderates for each other. This model could work well for when harassment comes in spikes of high volume (N3) so that moderator load is spread out.

Despite their feelings of guilt over burdening others, when we asked subjects whether they would moderate a friend's account, many were willing and even eager, with one person saying:

> "*I would be honored to do that for a close friend of mine or someone that I respect professionally, really any journalist that I was close to.*" [Jour1]

We additionally interviewed a close friend of Ex3, whom Ex3 said would be their chosen friend moderator. Ex3's moderator said:

> "*If I could help in any way, shape, or form, I would do that, no question... It's really difficult to watch someone that you care about so much go through this, and to be by-and-large helpless...to have a tool at my disposal that would help in even the smallest way, I would leap at a chance to do that.*" [Ex3]

Thus, though it is important to consider how to reduce moderator burden, we notice strong motivations for friends to help harassment recipients.

Reducing Secondary Trauma for Moderators

One concern with a friendsourced approach is whether it simply spreads trauma as opposed to reducing it. But when we asked subjects, they felt that it would be less traumatic for someone besides the intended recipient to read a harassing message, saying:

> "*I could emotionally handle reading someone else's hate if I'm far enough removed from it. It's not about you, it doesn't feel the same.*" [You3]

Ex3's moderator also felt that, as they do not personally know Ex3's harasser, the harasser would not be able to send targeted messages that would affect them. Despite the potentially lower impact that harassment could have on moderators, there is still risk of secondary trauma, as content moderators for platforms have described [28]. An idea subjects had for reducing secondary trauma was to choose moderators that did not share traits with the interviewee for which they would be harassed. One subject said:

> "*An army of woke cis white dudes would be great, because they're like, let's pay it back. Also, none of the harassment would be targeting their identity*" [You2].

This echoes work on the effectiveness of certain identities in bystander intervention [237]. However, Pub1 felt that certain insults targeted at an identity might not be recognized by people outside of that identity unless they were trained.

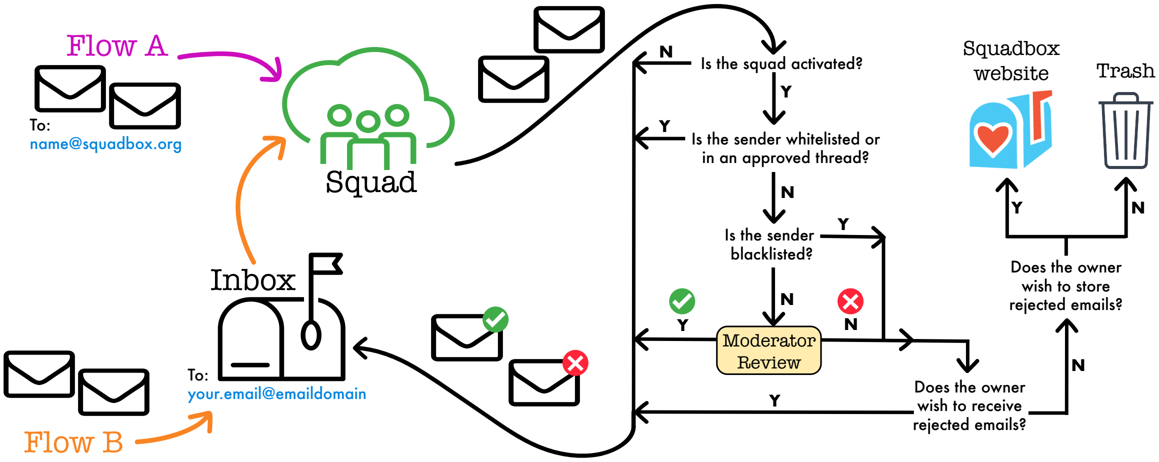

Flow A
To:
name@squadbox.org
Squad
Inbox
To:
your.email@emaildomain
Flow B
N
Is the squad activated?
Y
Y
Is the sender whitelisted or in an approved thread?
N
Is the sender blacklisted?
Y
N
Y
N
Moderator Review
Y
Does the owner wish to receive rejected emails?
N
Does the owner wish to store rejected emails?
Y
N
Squadbox website
Trash

Figure 6-1: Diagram of the flow of emails through Squadbox, including Flow A, which allows users to have a public moderated account, and Flow B, which allows users to get their current account moderated. From there, various settings define whether emails get moderated and where they go.

Summary

We determine several design goals necessary for a successful tool for friendsourced moderation. First, subjects described different preferences for what actions they wanted moderators to take and what powers moderators should have. Thus, any tool needs to be customizable to suit a variety of user needs and preferences (**G1**). Second, many subjects had messages they preferred to keep private, even from friends. While any such feature would already be an enhancement over the existing strategy of giving a friend one's password, a second goal is to allow users to mitigate privacy concerns (**G2**). Third, while subjects and their friends were eager to moderate, given recipients' guilt about asking for help and potentially high volume of messages, tools should effectively coordinate moderators and minimize their workload (**G3**). Finally, subjects expressed concerns about the emotional labor of moderators, motivating a final goal to minimize secondary trauma for moderators (**G4**).

6.4 Squadbox: A Friendsourced Moderation Tool

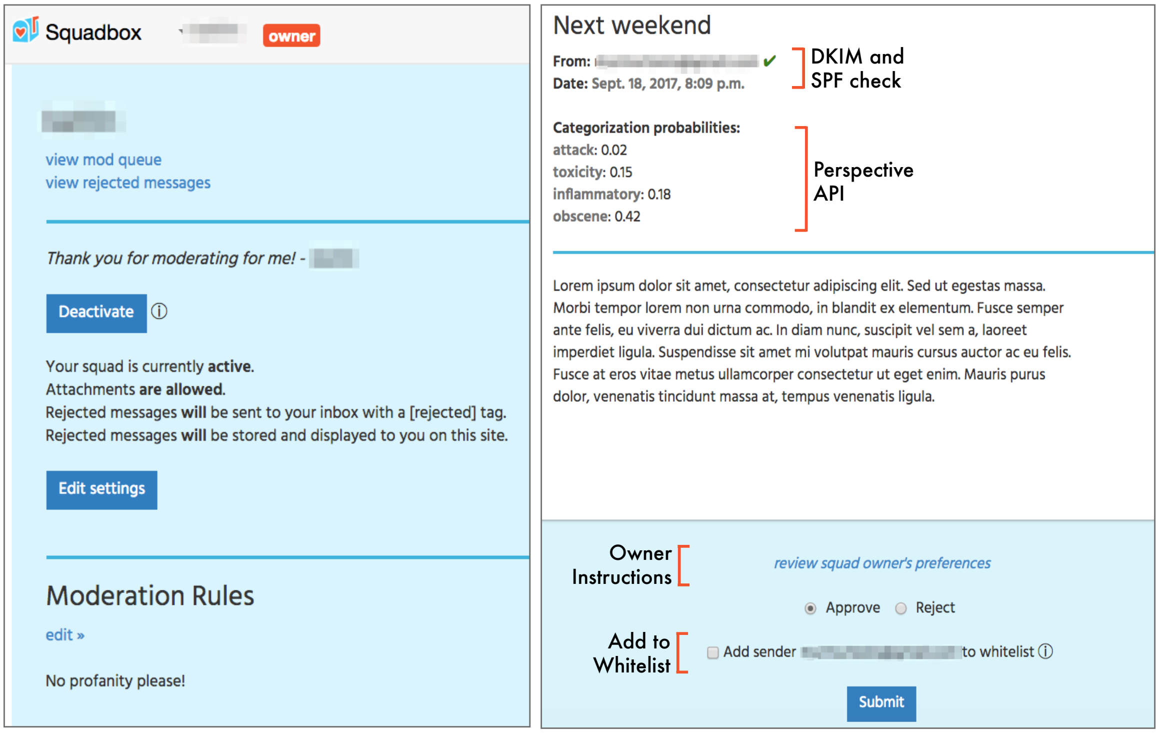
From the user needs and design goals arising from the interviews, we designed Squadbox[1], a system for recipients of harassment to have messages moderated by a "squad" of friends. Squadbox was developed for email as we discovered that email harassment was common among our subjects yet there were few resources for reporting harassment over email. However, Squadbox's general framework is applicable to any messaging or social media system, and we aim to extend it to them. We describe scenarios inspired by our subjects of how Squadbox can be used, with the workflow shown in Figure 6-1, followed by features and implementation of the system. From here onward, we use the term "owner" to refer to the person who is in charge of the inbox and having their emails moderated.

6.4.1 User Scenarios

Flow A: Squadbox as a public contact address. Adam is a journalist who gets harassment on Twitter due to his articles. He wants to have a publicly-shareable email address in order to receive tips from strangers, but is hesitant for fear of receiving harassment. Adam creates a Squadbox account, choosing `adam@squadbox.org`. He enlists two coworkers to be moderators because they understand context about him as well as his field. Adam uses his Squadbox account as a public email address. Any email he receives there goes through his squad first. In this way, Adam is able to open himself up to the public without risking further harassment (N2).

Flow B: Squadbox with an existing email account. The owner Eve is a professor. She has a publicly-listed email address through the university where she receives email from collaborators. Her research has been the subject of controversy, so she sometimes receives bursts of harassing emails. She wants to (and must) keep using this account for her work (N2), but cannot communicate when she's under an attack. Eve sets up a squad and asks her spouse and a friend to serve as her moderators. She sets up a whitelist and filters so that only strangers' emails go to Squadbox. She can

[1] Squadbox: `http://squadbox.org`



[redacted]

[redacted]

view mod queue
view rejected messages

Thank you for moderating for me! - [redacted]

Your squad is currently **active**.
Attachments **are allowed**.
Rejected messages **will** be sent to your inbox with a [rejected] tag.
Rejected messages **will** be stored and displayed to you on this site.

Moderation Rules

edit »

No profanity please!

Next weekend

From: [redacted] ✔
Date: Sept. 18, 2017, 8:09 p.m.

Categorization probabilities:
attack: 0.02
toxicity: 0.15
inflammatory: 0.18
obscene: 0.42

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed ut egestas massa. Morbi tempor lorem non urna commodo, in blandit ex elementum. Fusce semper ante felis, eu viverra dui dictum ac. In diam nunc, suscipit vel sem a, laoreet imperdiet ligula. Suspendisse sit amet mi volutpat mauris cursus auctor ac eu felis. Fusce at eros vitae metus ullamcorper consectetur ut eget enim. Mauris purus dolor, venenatis tincidunt massa at, tempus venenatis ligula.

review squad owner's preferences

● Approve ○ Reject

□ Add sender [redacted] to whitelist ⓘ

Figure 6-2: On the left, an owner's view of the information page for their squad. On the right, a moderation page for the moderator.

also turn on Squadbox when she starts getting harassment but then turn it off when it dies down (N3).

A second scenario for Flow B involves Julie, who is dealing with harassment from an ex-significant other. She cannot simply block this person because they need to coordinate the care of their child. Julie creates a squad of one close friend and sets up a filter to forward only emails from her harasser to her squad. Her moderator separates out and returns information about coordination while redacting harassing content (N4).

6.4.2 Squadbox Features

Now we turn to describing how Squadbox works for both owners and moderators, and how our features work to fulfill user needs and our system design goals.

Features for Reducing Moderator Load and Increasing Privacy

To begin, we describe automated moderation features that work to reduce the burden placed on moderators (G3) as well as support increased owner privacy (G2).

Filters: Squadbox supports filtering by sender whitelists and blacklists. We allow an unlimited number of email addresses to be whitelisted or blacklisted, meaning emails from those senders will be automatically approved or rejected, respectively, without needing moderation. We also allow owners to choose whether or not moderators can add to their whitelists or blacklists (N5, G1). Finally, we develop tools to easily import from one's contacts and export to filters. Such filters partially alleviate any concerns about slow moderation turnaround time, and helps owners feel more in control over what messages their moderators see (G2). There is significant room to expand this filtering capability by allowing owners to choose a specific behavior—approve, reject, or hold for moderation—for each message based on its content, sender's email domain, etc., or any combination of those.

Automatic Approval of Reply Messages: Owners can set Squadbox to automatically approve replies to a thread where the initial post was moderator-approved. We also allow owners to opt back in to moderation for a specific sender-thread pair. This feature provides more fine-grained control over how much of conversations moderators can see (G2), reduces the number of messages moderators must review (G3), and makes extended email conversations less hindered by the delays of moderation.

Activation and Deactivation: Several subjects mentioned periods of no harassment in between harassment, as well as times when they could anticipate receiving harassment (N3). To better accommodate this, users can deactivate a squad so that all emails will be automatically approved, reducing moderator workload (G3). When it is reactivated, all previously defined settings, whitelist, etc. take effect again.

Features for Reducing Secondary Trauma to Moderators

Now, we describe existing and planned Squadbox features that work to minimize secondary trauma to moderators (G4).

Control over Viewing Harassment: Subjects described how receiving harassment in their inbox disrupted their day-to-day (N1); similarly, receiving someone else's harassment in their inbox might disrupt a moderator. To prevent this, we only show messages on the Squadbox site, giving the moderators control over when to moderate. Extending this concept, we plan to protect moderators further by obfuscating all or part of image attachments and message contents and allowing moderators to reveal them as necessary. Machine learning models such as Perspective [111] could help determine what to obfuscate.

Limit Moderator Activity: When a new message comes in for moderation, we notify the least recently notified moderator, and only if they have not been notified in 24 hours. This makes it easier for moderators to step back from the task by limiting how frequently they are reminded of it. In the future, we aim to allow moderators to temporarily give themselves a break from seeing notifications or messages, allow owner- or moderator-set hard limits to moderation, and automatically check in on moderators occasionally. We also plan to publicize training and support resources for moderators.

Features for Giving Moderators Context and Information

Next, we describe features that give moderators more information to better tailor their decisions (G1) and make moderation easier (G3). These are shown in Figure 6-2.

Thread and Sender Context: Given that subjects said harassment is often repeated, having the context of a thread or all messages from a sender may help. Thus, we show the entire thread of messages to a moderator when they review a message. We plan to expand this by matching particular senders to particular moderators, or by allowing moderators to quickly review past moderated messages from a sender.

Customized Instructions: As people have different ideas about what is harassment [309] or have different actions they want moderators to take, we allow owners to give instructions to their moderators via a freeform text box (G1).

Verified Senders: We inform the moderator whether the message passes SPF and DKIM checking, which use cryptography to detect *spoofing*—senders pretending to be other senders to sneak past moderation. For senders that don't use DKIM or SPF, we implemented a simple hash-token system that allows senders to verify their identities via a secret shared between them and Squadbox. When they send emails to `squadname+hash@squadbox.org`, the email passes verification. A new hash can be generated if it gets compromised.

Automatic Harassment Signals: We provide machine-classified signals of messages' toxicity, how obscene or inflammatory they are, and how likely they are to be an attack based on scores provided by the Perspective API [111]. These scores are shown to moderators when they review messages.

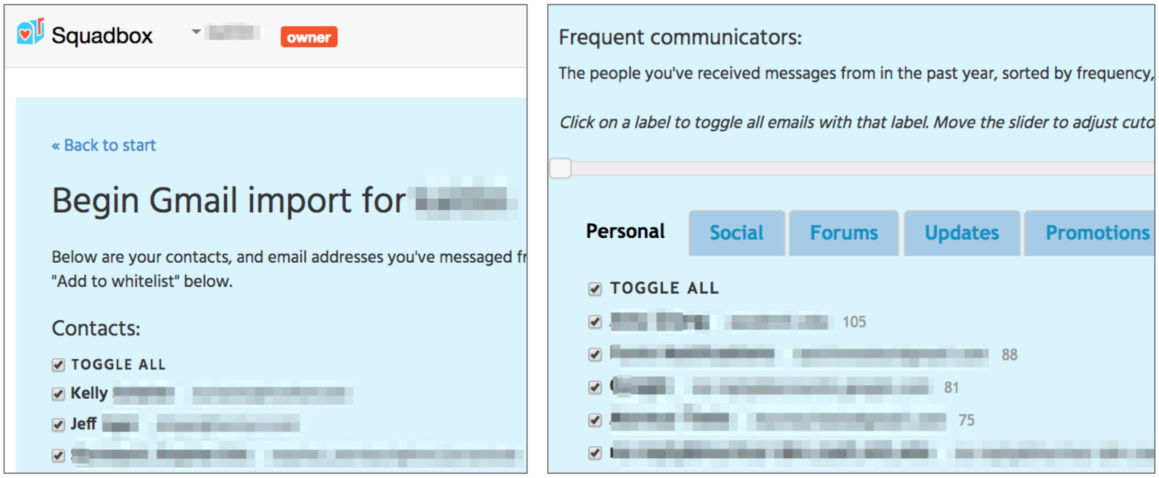
Features for Giving Owners Customization Capabilities

Finally, we describe features that allow owners to customize what should happen to harassing messages (G1).

Divert and Collect Harassing Content: We give owners the option to receive harassing content (N4) or file them into a separate folder (N1), given this request from interviews. Owners can choose to do one, both, or neither of the following: 1) receive rejected messages with a "rejected" tag, and 2) store rejected messages on the Squadbox website. We provide downloadable Gmail filters for owners to automatically forward emails with a "rejected" tag into a separate folder.

Moderator Tags: Several subjects said it would be useful to have their moderators add tags to messages, such as the nature of the harassment or its urgency. Currently, the moderation interface supports a list of tags indicating common reasons why a message might be rejected, such as "insult" or "profanity". If an owner has chosen to receive rejected emails, they are sent with the tags added in the subject line. Recipients can then add a filter in their mail client to customize where those messages go. They can also be grouped or sorted on the website (N6).

Moderator Explanations or Summaries: Some subjects thought it would be important to understand moderators' rationale for rejecting particular messages. Thus, we allow moderators to provide a brief explanation for their decision or a summary (N4). This is displayed in the web interface with the rejected message, and



Squadbox [redacted] owner

« Back to start

Begin Gmail import for [redacted]

Below are your contacts, and email addresses you've messaged fr
"Add to whitelist" below.

Contacts:

☑ TOGGLE ALL
☑ Kelly [redacted]
☑ Jeff [redacted]
☑ [redacted]

Frequent communicators:

The people you've received messages from in the past year, sorted by frequency,

Click on a label to toggle all emails with that label. Move the slider to adjust cuto

☑ TOGGLE ALL
☑ [redacted] 105
☑ [redacted] 88
☑ [redacted] 81
☑ [redacted] 75
☑ [redacted]

Figure 6-3: Squadbox generates whitelist suggestions from owner's Gmail contacts.

inserted at the top of the email if the owner has chosen to have rejected messages delivered.

6.4.3 System Implementation

Squadbox is a Django web application. Data is stored in a MySQL database and attachments in Amazon S3. It interfaces with a Postfix SMTP server using the Python Lamson library. We describe how the system works for both Flow A and Flow B, as well as optimizations for Flow B using Gmail.

Flow A: This flow works like a moderated mailing list with one member. Once messages have passed the moderation pipeline, we send them to the user's email address. If incoming messages are automatically approved by a filter, they are delivered immediately. Otherwise, they are stored on the server until they are moderated.

Flow B: This flow requires an extra step—we must first remove the message from the owner's inbox, and then potentially put it back. To accomplish this, the owner's email client must allow them to set a filter that only forwards some messages, for example, "forward messages that don't have `[address X]` in the `list-id` header field". We need this capability to prevent a forwarding loop—by slightly modifying messages that pass through Squadbox, we stop them from being re-forwarded to us. This capability is common in email clients (Gmail, Thunderbird, Apple Mail), but

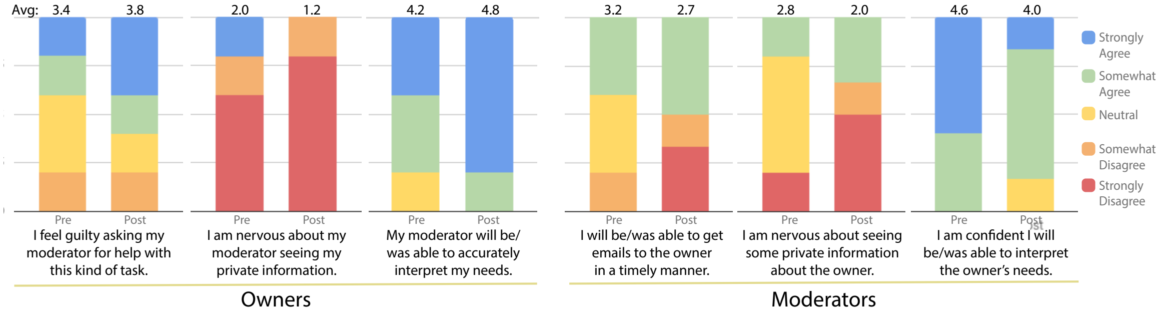


Figure 6-4: Comparison of agreement (where 1=Strongly Disagree, 5=Strongly Agree) with statements before and after the field study.

not universal. Messages from whitelisted senders or that are otherwise automatically approved are immediately sent back when Squadbox receives them; the rest are stored on the server until they're moderated. We provide instructions for setting up filters with the correct address. This address contains a secret hash to make it harder for attackers to send fake approved emails. However, if the address gets compromised, such as if the owner forwards an approved email to an unsafe sender, the user can generate a new address and filter.

For Gmail users, we leverage the API to add optimizations to mitigate privacy and security concerns and enhance the user experience. As in Figure 6-3, the owners' contacts are imported to generate whitelist suggestions. Gmail's rich filtering language allows us to generate filters to only forward emails needing moderation to Squadbox, giving owners greater control over which messages pass through the system. Accepted messages are recovered out of the trash rather than being re-delivered via SMTP, meaning the recipient sees the original message.

6.5 Evaluation

Due to the sensitive nature of online harassment and the uniquely vulnerable position of its recipients, we were wary of conducting a lab or field study with recipients of harassment for fear of potential negative consequences for participants. For owners, we worried that if anything were to go awry (for example, lost emails) we would be causing further damage to an already vulnerable group. For the owners and even for moderators, there may be psychological risks to reading harassment (either real, or even simulated for the purpose of a study). We also feared that persistent harassers could become aware subjects were using Squadbox, and seek out security vulnerabilities. All of these concerns compel us to take the necessary time to convert our research implementation into a full-fledged production system before actual usage trials. In preparation for an initial launch, we presented a demo of both the owner setup and the moderator workflow over screenshare to five of our interview subjects. Additionally, in the interest of evaluating the usability of our system and further contextualizing friendsourced moderation, we conducted a field study with five pairs of friends, where the owner was instructed to have moderated any emails they did not wish to receive. For our test subjects, this was mostly spam and advertisements.

6.5.1 Feedback from Demos to Harassment Recipients

We demoed and discussed the Squadbox tool with five of our interview subjects, Pub1, Res2, Ex3, Act1, and Act2, for 30-40 minutes to get their feedback on the possible settings and the workflow. All the subjects indicated that Squadbox's settings were flexible enough to capture the way *they* would want their email handled. Asked about willingness to let their email flow through Squadbox, all subjects were comfortable with the level of access that Squadbox required, and expressed interest or even excitement to use the tool, with Pub1 saying:

> "*I would tell you this is a very strong pragmatic tool...Overall I think it's in really great shape [to make] a beta and I'm very excited about this.*"

Subjects also had ideas for further customizations, such as the ability to create template responses for moderators to send back to people, modules to train new moderators about specific identity-related attacks, and obscuring sender email addresses (which can themselves contain words that harass). Three subjects were concerned about design aspects that would make it too easy to go read their harassing emails out of curiosity. They wanted ways to make it harder to see that content, such as requiring the owner to ask their moderator for access. One subject wanted sender

Squad	WL Size	% Accept	% Reject	Total Vol.
S1	231	32	68	22
S2	333	44	56	77
S3	929	32	68	37
S4	19	29	71	139
S5	122	100	0	25
Avg.	326.8	47.4	52.6	60

Table 6.2: Usage statistics by squad. Whitelist size, followed by percentages of messages approved and rejected by the moderator during the study, and a total count of all manually moderated messages.

identity obfuscation, for fear that moderators may try to retaliate against harassers.

6.5.2 Field Study Methodology

We conducted a four-day field study with five pairs of friends (three male, eight female, average age 24), where owners were recruited via social channels, and they were asked to find a friend moderator. Owners were required to use Gmail, while moderators could use any email client. One owner chose to add a second friend moderator during the study. To begin, we helped owners set up their Squadbox account, whitelist, and Gmail filters either in-person or over video chat. Once their friend accepted a moderator invitation, we explained the workflow to moderators over email. Moderators were asked to moderate emails for the owner at their own pace throughout the four days. At the end of this process, we asked both owner and moderator to complete a survey about their perceptions of the tool and friendsourced moderation.

6.5.3 Field Study Results

The whitelist/blacklist feature was an effective way to separate out potentially unwanted messages. As shown in Table 6.2, in all but one squad, the majority of messages (52.6% overall) sent to moderation were rejected. This suggests that whitelists, along with the automatic approval of reply messages, worked fairly well to avoid moderating emails users did want. For the squad (S5) where that was not the case, the owner's rules were extremely limited, while the other owners had given more specific instructions; for example:

> "*I don't want emails from all those job companies or from student organizations from my previous schools. Research group-related emails are fine.*"

Future work can optimize this even more using richer filters or human-in-the-loop machine learning.

Both owners and moderators relied on outside knowledge and communication about the owners' preferences. Although we asked owners to write moderation rules, these were all rather short (2 sentences or fewer). Owners hoped their moderators would understand what they wanted:

> "*I felt like I was putting a lot of trust in [my moderator] knowing a lot about me.*"

At the start of the study, moderators said that outside communication would be useful to them for clarifying what owners wanted:

> "*I am a bit concerned but I know that I can clarify with her whenever there is a need. I will ask her because I am in constant contact with her.*"

Both owners and moderators noted after the study that they used this strategy to resolve uncertainty. A moderator said:

> "*There was some ambiguity at the beginning, I contacted the owner and she clarified it for me.*"

And an owner stated:

> "*We talked about certain messages and determined whether to add the sender to the whitelist.*"

Owners and moderators became less concerned with privacy over time. As shown in Figure 6-4, both owners' and moderators' concerns about privacy decreased about the same amount during the study. Interestingly, moderators were overall more concerned with privacy than owners. This may be because owners went through the whitelist process and thus were more confident that they would not forward private information, while moderators had no knowledge of what owners were forwarding or not forwarding.

Both owners and moderators became less likely to think messages were handled in a timely manner. Both groups decreased in their confidence in timely delivery. Additionally, after the study moderators said on average that "moderating is a lot of work". One owner added a second moderator during the study because the first one was busy for one of the days. Although a majority of decisions led to "reject", we did not see active use of the blacklist feature, suggesting that it may be important to allow the creation of more fine-grained blacklist rules, such as ones containing both an address and phrase.

While owners grew more confident in their moderators over time, moderators grew less confident in their own abilities. This opposite change between owners and moderators can be seen in the third and sixth statement in Figure 6-4. In addition, owners felt more guilty over the study.

6.6 Discussion

The field study suggests that, despite a close relationship and open communication between owners and moderators, tensions may still arise around timeliness of message delivery, moderator burden and guilt, and perceived performance. These tensions may arise because friends are performing a favor to the owner, so owners feel both grateful but also guilty about the exchange, and decline to voice concerns about timeliness. Conversely, a friend may feel the burden of responsibility towards the owner and worry that they are not doing enough. Some of these issues might be addressed with additional feedback in the system, such as allowing owners to show appreciation, or for moderators to be able to communicate when they will be unavailable. Concerns about timeliness also stress the importance of having multiple moderators. Another approach could be "soft" moderation, where thresholds for moderation vary dynamically to limit moderators' workloads. The field study also showed that concern about privacy was overall minimal and that moderators were able to infer owners' desires or ask for clarification.

Finally, we noticed that owners had widely differing settings for their squads, using them to tailor moderator privileges and automatic rules to their liking.

6.6.1 Friendsourced vs. Volunteer vs. Stranger Moderation

While most of our interviewees and field study subjects preferred friendsourced moderation, a few YouTube subjects and Pub1 were more interested in paid stranger moderators because they considered their activity a business and did not wish to exploit friends' unpaid labor for it. However, these interviewees felt it would be important for the moderators to be vetted, trained, and have established trust. This suggests that the approach of prior systems such as EmailValet [184] may not be appropriate. We note that, despite their interest, You3 and You4 stated this would not be financially possible for them. This suggests that there may be room for innovation in a moderation tool that has lower costs at scale but still provides some assurances of privacy and quality. One subject, Pub1, did pay moderators but gave them direct access to their account, causing privacy concerns. Pub1 described their workflow as "cobbled together", and expressed enthusiasm about Squadbox making moderation easier and about whitelists for improving privacy. A final population is volunteer moderators, much like the vetted community within HeartMob [26]. However, we would need to set checks to protect against harassers seeking to infiltrate the system.

6.6.2 Harassment on Different Platforms

The present-day siloing of online communication into numerous platforms is a boon to harassers, as harassment protections must be designed and implemented separately for each platform. As we saw in interviews, recipients are often harassed on multiple platforms at once. Indeed, because some harassers are determined, if one platform becomes more adept at dealing with harassment, recipients may start receiving more harassment on other platforms. This is why some subjects did not want harassers to know that they would be getting their emails moderated, as this might just increase their harassment elsewhere. But if Squadbox or a similar tool succeeds in becoming popular, then simply trying to obfuscate its use would likely fail. As a result, harassment recipients are as vulnerable as the "weakest link" in their suite of communication tools. To combat this problem, we would like to expand the capabilities of Squadbox beyond email, to other encompass other platforms. However, we must rely on and build for each platform's API, and develop browser extensions or native clients. A far better solution in the long term would be to evolve a single, standard API for accessing messaging platforms. After all, whatever extra features they provide, each platform's model is at its core just a collection of messages. Given such a standard API, a single tool could tackle harassment on all the platforms simultaneously. Unfortunately, such an API seems inimical to the business model of these platforms, as it would enable users to access their messages through third party tools and avoid visiting the sites at all.

6.7 Limitations and Future Work

In our implementation of Squadbox, we encountered some issues with rate-limiting in the Gmail API, as well as issues where emails from domains with strict DMARC settings were rejected by email clients. IMAP is currently implemented using mailing list APIs, but in the future we plan to re-implement Squadbox as an IMAP client, giving it more power to fetch email from any IMAP server and easily move email between folders using the IMAP protocol. Since multiple clients can access the same server, owners could still use whichever email client they prefer. Additionally, despite the limitations described in the previous section, we plan to connect Squadbox to other communication platforms. Finally, while our field study explored the use of Squadbox as a friend-moderation tool for email, it did not study recipients of harassment. Of course, there are many differences between spammers and harassers, including that harassers are often much more determined when targeting a particular person than spammers, and that the content that harassers produce has an emotional toll. There are also still many potential security issues to address, such as fighting email tracking techniques [80]. In the future, we aim to move cautiously towards releasing Squadbox, including giving more demos to harassment recipients and their potential moderators before initiating a small-scale release.

6.8 Conclusion

In this work, we study the emergent practices of recipients of online harassment, finding from 18 interviews that many harassment recipients rely on friends and family to shield themselves from harassing messages. Building on this strategy, we propose friendsourced moderation as a promising technique for anti-harassment tools. We developed Squadbox, a tool to help harassment recipients coordinate a squad of friends to moderate aspects of their email. From a field study, we found that the use of friends as moderators simplifies issues surrounding privacy and personalization but also presents challenges for relationship maintenance.

Chapter 7

Discussion

The four systems that I described demonstrate the possibilities for *collective discussion curation* online. Gathering from our empirical needfinding data, design explorations, and observations of and interviews with people using the tools, I compile a series of design implications for the representation of curated discussion artifacts. Following this section, I discuss how this work fits into a broader framework of discussion curation tools, considering both the possible scope of curation decisions and the degree of automation versus human input within curation tools. Finally, I turn towards what barriers still exist towards making collective discussion curation a part of our everyday online discussion tools and describe the normative changes that still need to occur on the part of users and system operators.

7.1 Design Implications for Discussion Curation Artifacts

In this thesis, I describe empirical evidence regarding the kinds of information representations and interactions users would like as well as proposed and tested a number of novel designs for discussion artifacts arising from curation. This data-gathering informs the following implications for designing discussion curation artifacts.

7.1.1 Superimposed Structure

From our user study of Wikum, we found that people were reluctant to edit each other's work, something that has also been observed in other studies [11]. We also found in our studies of both Wikum and Tilda that users wanted the ability to read the raw discussion in the authors' original voices. This was echoed in our interviews with Wikipedia RfC closers who wanted a way to easily see original discussions even if someone they trusted had summarized some or all of the comments.

From these two takeaways, I present the design implication of supporting *superimposed structure*, where any curation performed on a discussion is overlaid on top of that discussion as opposed to editing or destroying it. As a result, curators would have fewer misgivings about transforming another person's statements since the original would still be there. Superimposed structure also means that curators need not be tied to one particular structure for an artifact and can add structure over time. Hesitancy to commit to a particular structure prematurely is one reason why people resist formalization [300]. Indeed, if it was deemed necessary, one could imagine superimposing different structures on top of the same original discussion for different use cases. However, one challenge of this approach is determining how to represent conflicting structures, such as when two people have different ideas about how something should be summarized. One option is to allow the system to branch and users can select which branch they prefer, though this could get complicated.

7.1.2 Fine-grained Hyper-linking

For the user, superimposition requires a way to travel from the curated artifact to the relevant original content. This is achievable using *hyper-linking* between the two. Not only is it important to link from curated material to original, links also need to be *fine-grained.* Given that we are dealing with large discussions with as many as thousands of comments, it would be useless to have a single link to the original content. Instead, each portion of the curated artifact that deals with a particular sentence, comment, or small set of comments should be able to link directly to that content.

This is beneficial for several reasons. First, more fine-grained hyper-linking from a curation artifact makes navigation of the original discussion easier. Even if a reader wanted to read the discussion in its entirety, they could still use the curated artifact to choose where and in what order to dive in [266]. For example, we saw in Tilda that people wanted to go from a line in a summary to its corresponding original chat message so that they could read the chat message in context. The summaries became a portal into an unorganized and indistinguishable stream of messages. For anyone who mistrusts what a curator has done, they can verify the curator's work by checking the relevant original content directly from the curated artifact.

In addition, the process of building these links encourages curators to "cite their work", a practice that can both help curators develop a more neutral point of view [224] as well as signal greater credibility to their work for readers [384].

7.1.3 Iterative Curation and Hierarchical Navigation

Curating an information artifact is generally a process of distillation, which necessarily removes some information while hopefully keeping the most salient parts. Distilling a discussion that is thousands of comments long to something short enough to be reasonably readable necessarily loses a lot of information. The process of creating such a large jump in condensing of information can be cognitively challenging, as one needs to juggle all that content in one's head and also make many difficult decisions about what to include. We saw when talking to Wikipedia RfC closers that they would spend up to 4 to 5 hours both reading and drafting a closing statement and often felt it was necessary to do the work in one sitting so they would not forget what they had read.

Instead of requiring just one level of distillation, systems like Wikum allow curators to create as many levels of summaries that they want, in an *iterative way*. This lets off some of the pressure for curators as they can consider smaller portions of the discussion at a time and also work their way up towards a short summary. The final summary tree artifact in Wikum then permits *hierarchical navigation* so that readers can choose whether they want to read at a higher level or go more in depth. Similarly, moderators in Squadbox can iteratively mark emails before deciding to add an email account to a whitelist or blacklist.

While hyper-linking in the context that we've used it refers to document-based modes of hypertext, where nodes focus reader attention and links are for traversal between nodes, we can also incorporate notions of *spatial hypertext* on the page [219]. That is, the size, shape, and placement of nodes can signal their structure and context. Spatial hypertext is best suited for cases involving more exploratory structuring where readers and writers are the same. We use spatial hypertext in the visual representation of the summary tree in the Wikum interface and presentation of Tilda summaries.

7.1.4 Discussion Curation Artifacts as Boundary Objects

Finally, we saw in our interviews with discussion participants and curators that people wanted a way to reference or get notices about discussions that crossed different platforms. For instance, in interviews with people who use both group chat and email heavily for work, users described losing track of discussions that happened in both places, with no way to bridge the two mediums. Tilda users as well as Wikipedia closers also wanted the ability to take discussion summaries and propagate them to other places. Concepts from tools for curating workflows like Murmur could be used to design ways people could transform and repost discussion curation artifacts. For instance, Tilda users wanted a way to take action items and decisions marked in Tilda to appear in task management systems such as Trello. Squadbox users wanted a way to share and remix blocklists.

Thus, discussion curation artifacts should be designed to serve as boundary objects that can be passed around between different communities of practice [314], such as from a workteam to managers, or from a group chat channel to an email thread, or posted to a task management system. The ability to add clarifying metadata to artifacts allows them to serve as a suitable translation between communities and discussion environments. Another criteria is to preserve linking between the different places where the artifact resides, so that users can always trace back to the original discussion. Given links, it should then become possible for artifacts to allow *dynamic updating*, where changes made in one place can flow to the other sites. Design decisions such as keeping track of edit history and making visible curators and the aggregation of their contributions would help to promote trust in the artifacts and provide both accountability and credit to curators' work.

7.2 Considering the Scope of User Curation

I now turn towards placing the systems in this thesis into broader frameworks of discussion curation tools. One way in which the four systems differ from each other is in the *scope* of curatorial actions. Each system involves a set of curation actions; sometimes these actions affect only one individual, while other times they affect everyone using the system. When should actions affect everyone versus only the person making the action? In this section, I discuss the common scopes, their pros and cons, and finally argue for more tools that enable a *networked scope*, which can grow and shrink as needed according to social proximity.

7.2.1 The Technical Scope of Curation: Platform, Group, Network, and User

In the case of Wikum, users are contributing to the same shared artifact with no branching or personalization of what each user sees, much like a Google Doc document. In the case of Tilda, again, each user's actions reflect on a shared artifact. Tilda allows users to personally subscribe to different summaries but the summaries themselves look the same to everyone within the Slack group. For both Wikum and Tilda, the scope of curation is **group-wide**—each user action affects everyone within a group using the system.

In contrast, Murmur makes it more possible for everyone to see something different. This is because the curations that users can perform in Murmur primarily alter only their own personal settings as opposed to system-wide settings. For instance, I

Scope	**Definition**	**Thesis Examples**	**Other Examples**
Platform	Curation actions affect anyone using the platform.		Platform-wide algorithms for sorting or filtering, platform-wide content moderation decisions
Group	Curation actions affect everyone in the group.	Wikum, Tilda	Wikipedia or StackOverflow editing, user bans or content takedowns by a forum moderator
Network	Curation actions affect only those who have given permission, directly or indirectly, to the user making the action.	Squadbox	Twitter blocklists, shared email filters
User	Curation actions affect only the user making the action.	Murmur	Starring or bookmarking an item, muting, blocking, or following a user, personal filters, personal inbox folders

Table 7.1: A framework of curation broken down by their technical scope.

can choose to subscribe to only emails about apartment listings but this doesn't affect the emails that other people receive. Similarly, blocking a user is a personal action that I can take that affects no one else. Thus, the scope of curation is at the **user** level. Exceptions include the tags that users can add to emails that then become shared throughout the group.

Unlike the other tools, Squadbox permits users to curate content for specific other users. Not everybody can curate content for everybody else, like in the group-wide case. Only once someone has granted permission to someone else can that person come in and curate for the person who gave them permission. In this case, the scope of curation is **networked**—a user's curation actions affects only those who have chosen to accept that user's action.

Finally, one scope not explored in this thesis is at the **platform** level. These curations affect anyone using the platform, regardless of what groups they belong to within the platform. For instance, many of the algorithms that dictate sorting or filtering of content on major social platforms operate platform-wide.

These four categories can characterize many of the scopes of curation actions technically available within online discussion systems. Of course, most sociotechnical systems incorporate a range of curation powers at varying levels of scope. For instance, in Tilda, curation actions to generate summaries are at the group level but personally subscribing to a particular set of summaries is a user-level action. There are also other ways to interpret "scope" beyond what is *technically* possible. For instance, in a normative interpretation, one user's decision to block someone may encourage others to do the same. In the words of Lessig's *pathetic dot theory* [205], I am focused on the force of architecture or technical infrastructure, as opposed to law, norms, or markets.

More examples of where different curation tools sit can be found in Table 7.1. As can be seen, quite a lot of curatorial actions fall at the platform, group, or the user level. While systems like Facebook and Twitter have popularized the concept of networked *sharing* of content, there are few abilities to have networked forms of *curation*. One well-known example of networked curation is Twitter blocklists [143, 102], where users can collaborate on shared lists of accounts to block, and they can also subscribe to each others' blocklists. This feature is not actually built into the Twitter platform but developed by volunteer third parties seeking to combat rampant harassment on Twitter. Though less common today, email users in some systems can similarly share spam filters with each other in a peer-to-peer way [116].

7.2.2 Localizing Curation

One way to think about the scope of curation actions is as a form of localization or decentralization of governance. Researchers have argued that for democratic governance to succeed in offline governments, communities need strong local civic associations to generate social capital [273]. In the online setting, there may not necessarily be geographic constraints to determine what is sufficiently local, particularly when the "group" is large, perhaps as large as the platform itself. In these cases, the strength of social ties within an online community could help determine the bounds of more "localized" decisions. This is the potential behind networked curation, which can flex to cover curation decisions that need to operate somewhere between group-level and user-level. In the example of Squadbox, users explicitly grant powers to specific other users to curate for them. One could also more nuanced forms of delegation or propagation of curation actions based on an underlying social landscape.

Given the relatively few active examples, I argue that *we need more networked curation tools.* What is the benefit of having curation scoped at this level? First, as mentioned above, networked curation is decentralized like user-level curation, which means that decisions can be tailored to localized needs as opposed to needing to suit an entire group or platform. However, unlike fully individual tools, users can share the load of any necessary manual work, which can oftentimes be too much for an individual to bear, as we found in the case of online harassment. The sharing and remixing of curation artifacts would also allow more difficult or cumbersome forms of user curation to scale more easily.

However, this doesn't mean that all curation should be scoped to the user and network level. One downside of more decentralized curation within a group is the resulting reduction in common ground between group members, a key ingredient to productive collaboration [249]. It may also reduce feelings of common identity or organizational commitment towards the community [279]. For example, the forum software Discourse [71] adds features to try and encourage unity when users are in conflict. While they allow users to mute each other, users must set an expiration date with the maximum value set to four months. In addition, once a user has been muted by five people, system administrators are alerted so they can consider whether to ban or intervene with the user rather than splinter the community.

Loss of organizational identity is certainly a major pitfall that communities should be concerned about. However, some social spaces such as Facebook today are so large that there is no real sense of community or social identity as a whole. Particularly in these cases there may be fewer downsides to loosening requirements for curatorial consistency and allowing more decentralized forms of curation to flourish.

Finally, it seems clear that some forms of curation should be scoped more broadly while others need more local differentiation. Take the case of online harassment—

tactics such as “doxing” or posting of non-consensual sexual imagery can only be addressed at the platform level as harm is inflicted by that content if it appears to *anyone*. On the other hand, we found that users have very different ideas of what they considered harassment when it was targeted at them, so a networked or user-level scope would allow people to take their own context into consideration. More work is needed to understand how to consider scope in the design of curation tools for particular cases.

7.3 The Role of End Users in Discussion Curation

In this thesis, much of the focus was on tools for *end user* curation, including ways to distribute and aggregate the work of end users and ways to give end users more expressive powers, so that they might bring their insight and knowledge of context into curation. In addition to end user curation, there are automated methods of curation as well as human curation by teams of moderators. Earlier in Section 1.2, I argued that many of the problems that online discussion systems face cannot be addressed by the current capabilities of automation or centralized forms of human moderation. However, automation has its benefits—namely, the ability to make many decisions quickly and cheaply—a useful trait when we are considering problems of information overload. And there are some upsides to end users doing less curation in certain cases, such as reducing end user burden and delegating decisions to people with greater expertise or training.

In this section, I argue that there are two types of issues when it comes to human involvement in curation systems. The first deals with *how* humans are involved, where the problematic cases stem from *incorrect emulation of user preferences*, or where the curation done on behalf of a user is failing to do what that user would have done. The second type is *which* humans are involved, where problems arise due to a *lack of representation*, or where the people who are designing, deploying, and wielding curation tools do not represent the people affected by their actions.

How Are Humans Involved	Which Humans Involved	Examples
Case-by-case human judgment	End users	Collaborative authoring (Wikipedia), annotation, note-taking, starring, tagging, bookmarking
	Community leaders	Community moderation, intervention (subreddit or Facebook Group moderators)
	Paid moderators	Commercial content moderation: Facebook, Twitter, etc.
Algorithms hard-coded by humans	End users	Email filters, If This Then That [153], blocklists, mutelists, followlists, word filters
	Community leaders	Wikipedia editing bots, Reddit AutoModerator
	Platform developers	Sorting (Reddit or Facebook comment ordering), some forms of filtering, clustering, trending algorithms
Algorithms learned from humans	End users	Algorithms trained on user input (Slack Highlights [307]), personalized algorithms (Facebook newsfeed sort)
	Paid annotators	Classifiers (hate speech detection, sentiment analysis), multi-class classifiers (Gmail folders)

Table 7.2: Categories of curation systems according to how and which humans are involved.

7.3.1 Human Involvement in Curation

To begin, I first outline different curation systems separated out by how humans are involved and which humans provide input, as shown in Table 7.2. Grimmelmann makes a distinction between two modes of automatic versus manual moderation [119] but there are more gradations between the two extremes.

First, we have systems that involve **case-by-case human judgment**. Despite not using any algorithmic curation, these systems can scale by incorporating more people to do the curation work. One way is through greater decentralization, where everyone curates at the scope of the user, as we have seen in this thesis. This by nature takes into account each user's insights. However, an end user who is overwhelmed may have difficulty scaling up their personal curation work. There are also curation systems where community leaders make decisions, such as moderators in a subreddit or Facebook Group. While these leaders are members themselves and somewhat accountable to members, they also can be overwhelmed by the work of curation (sometimes called "mod burnout") [294]. Finally, there are human curation systems conducted by paid moderators, termed "commercial content moderation" [281]. These assembly-line systems operate at scale in many for-profit social platforms, where moderators are tasked with enforcing a standard set of guidelines. Moderators are generally hidden from users, may not be demographically representative, and oftentimes are not even in the same country as the users for whom they are moderating. As a result, users have described these systems as opaque and frustratingly inconsistent [107].

The next category involves **algorithms hard-coded by humans** that can then be applied automatically to any number of decisions. The creators and operators of these algorithms can be end users in the case of systems like filter lists and small conditional statements [153] that require no programming expertise. Operated by end users, these filter lists encode contextual information but are brittle and require constant supervision, as we saw in the case of online harassment word filters. Operators can also be community leaders who maintain group filter lists or activate small programs like an editing bot on Wikipedia or Reddit's AutoModerator on behalf of the community. Finally, hard-coded algorithms can be put in place by platform developers towards tasks like sorting or filtering comments on a page. These algorithms are sometimes generic algorithms designed by software engineers or mathematicians that can apply to many types of problems. Other times they are designed for a particular task, and any weights are hard-coded in by platform developers. End users generally have no say when it comes to the design or use of these algorithms.

Finally, there are curation systems involving **algorithms learned from humans** that then mimic their actions. The humans involved could be end users that are performing actions as part of any everyday activity. This allows an algorithm to be personalized to a user; for instance, the Facebook newsfeed algorithm that curates what a user sees is partially trained on that user's past actions. However, algorithms that learn from human actions don't necessarily know *why* the humans make those actions. In some cases, users may not want curation based off of their prior actions; for instance, people may click on clickbait but not want to see more of it. The humans involved could also be paid annotators who have been taught how to label a dataset of content. Software engineers then take the data, analyze and clean it, and train a model based on it. However, the paid annotators may not be representative of the actual users of these models. In addition, the models may need to be continuously updated with new training data, as user interpretations change or as new forms of content arise.

7.3.2 When Curation Tools Incorrectly Emulate Users

When a curation system emulates users' preferences incorrectly, this can be due to *random errors* in the system or *systematic biases* encoded in the system or process.

In the case of manual curation, errors leading to low precision are simply mistakes made by users or paid moderators. Human error tends to be rarer than algorithm error, as both hard-coded or learned algorithms for the most part have yet to reach human performance on curation tasks. The precision of different algorithms can greatly differ however. As mentioned, hard-coded algorithms such as filter lists can be brittle, particularly in the face of an adversarial actor. Thus while filter lists make use of more automation than a fully manual approach, they still do not always scale well. Other hard-coded algorithms, though not as brittle, still may need regular attention as people try to game the algorithm by learning its weights. One example is algorithms for trending topics or hashtags that must handle coordinated inauthentic activity. Learned models can also be error-prone due to an inability to understand the complexity of language [147].

Curation systems can also incorrectly emulate users due to biases, or *systematic* errors, as opposed to one-off errors. This can happen in manual curation if the community leaders or paid moderators are overall biased in a way that disagrees with end users. Better training or collective action on the part of users can help rectify this gap. Biases can also creep in to algorithms, even if end users are the ones providing input. For instance, a user-authored word filter to catch hate speech might systematically catch reclaimed uses of a slur while failing to catch more veiled forms of hate speech because of the rudimentary nature of the algorithm. Learned algorithms may also become biased because humans encode their own biases into the data that the algorithms learn from [22]. For instance, users are cognitively biased towards shocking or polarizing content, and thus sorting algorithms learning this behavior will also bias in that direction. Finally, learned models represent a static snapshot of human interpretation of content but these interpretations may be continually changing over time, leading to a recency bias.

One solution to these problems is to just improve algorithms by building more precise models to reduce precision errors and reweighing to reduce systematic errors. Other solutions involve collecting different data, such as collecting more contextual information to inform a decision or more representative data to reduce a bias. For example, learned algorithms could be trained on what users *want* for curation as opposed to what users *do*—this might require users to actually conduct more curation work towards training models however. While there are many efforts in both these direction, algorithms still fall significantly short when it comes to emulating users' needs. As a result, user-led curation will continue to be essential moving forward. Users can better scale up their manual efforts with tools that permit more collaboration, sharing, and remixing, such as the ones described in this thesis. In addition, tools could be developed that give end users the ability to build more sophisticated models than filter lists.

7.3.3 Giving End Users Representation in Curation Decisions

Finally, problems can arise when the people making decisions about curation are not the same as the people affected by curation. This *lack of representation* leads to an illegitimate propagation of a particular set of human insights and values. For instance, algorithms encode preferences through some form of learning from data generated by humans or hand-coding by humans. If users do not have a hand in deciding what goes into these algorithms, they have no say in the preferences that power their curation. The same is true for some forms of manual moderation, such as commercial content moderation, where users have little to no say in the process.

Users can feel unrepresented because they *cannot personally contribute* data or hand-coding to an algorithm, or cannot personally participate as a human moderator. Users may also feel unrepresented because they have *no representative* in the curation process that is advocating for their perspective or that knows their complaints. Finally, lack of representation can also occur because the target population is *too diverse* to be well represented by any one process. Some populations are so large that this is bound to be true, such as in the case of algorithms deployed on large social platforms with millions or billions of users.

One solution to a lack of representation is to give end users more powers to conduct or participate in curation so they can *contribute their individual voice.* What does this look like in practice? Manual curation or hard-coded filter lists by end users allow them to contribute their own perspective, and this thesis strengthens these efforts. In the case of platform-wide manual moderation, giving end users more direct voice in processes could involve methods of direct or participatory democracy, such as voting or citizen juries. In one of my recent projects, we are considering applying a constitutional jury system to platform content moderation.

When it comes to algorithmic systems, such as learned algorithms, end users could still provide their individual voice by having a say in data collection. For instance, tools could allow end users to choose what kinds of data and whose data goes into the algorithms they wish to use. Systems could also allow end users to pool their data to develop algorithms that suit their collective needs. Also, while it may be more difficult for end users to *build* algorithms, it should still be within their abilities to be able to decide whether and when to use what algorithms are available.

A second solution to a lack of representation is to give end users the power to *designate representatives in the process of curation.* Various democratic governance structures could be used, including elected representative democracy or liquid democracy. These representative methods could be used towards algorithm development that then encode the values of different stakeholders [386]. For instance, in Wikipedia, a Bot Approvals Group staffed by volunteers in the editor community oversees the approval of new bots that operate on Wikipedia. Group members must publicly nominate themselves and pass a public deliberation process to be approved to join the group.

A first step towards achieving representation is a basic level of awareness and education on the part of end users. This involves understanding what curation practices govern users' online social spaces. There also needs to be transparency into what kinds of features or constraints are encoded into algorithms, how any data was collected, and the activity and performance of the algorithm over time as it is deployed. Beyond that, users should be able to collectively petition or appeal in support of changes in a way that invites accountability.

7.4 Normative Changes

Having presented empirical evidence that collective discussion curation is feasible, fulfills a need, and is beneficial to users, I now turn to what barriers exist to making this vision a reality in our online social landscape and what aspects of design still need to be explored.

The systems that I developed are *sociotechnical* in that they combine both technology and people, and the social parts of each system needs to work in tandem with the technical parts. In each of the four systems of this thesis, I sought to work with particular communities, understand their needs, and then develop systems that could fit into their existing social environments so that they need not make large adjustments in their regular workflows. However, all of the systems do involve some shift in users' and platform operators' perceptions, including their outlook on what an online discussion system is for, what is the role of users, and what is the role of platform operators. While engendering normative change was not the focus of this thesis, it is still an essential aspect of improving online discussion, and could potentially be nudged by system design.

7.4.1 The Goals of an Online Discussion System

Many major discussion systems today are private corporations that provide their tools and services to users free of charge in exchange for their data and attention, which they sell to advertisers. The incentives given this arrangement on the part of platforms are to collect more data by encouraging users to author more content about themselves and each other, and to maximize engagement by encouraging and promoting highly engaging content. As mentioned at the beginning of this thesis, these goals are fundamentally misaligned with the goals of a healthy, sustainable community. When pushed to execute on their goals, social platforms invariably make decisions, such as introducing algorithmically sorted newsfeeds to surface highly engaging content, that reduce users' power to control what they see.

After the revelation of the deep problems plaguing online discussion systems followed by intense media scrutiny and public outcry, many platforms within the last year have chosen to orient their goals more towards promoting "meaningful interactions" [234] or "healthy conversation" [134]. However, this does not go far enough, as it is still platforms themselves defining what makes an interaction meaningful. Instead, the goals of an online discussion system should instead be to support individuals and communities of interest in defining their own discussion space. In the words of Lessig, platforms could move from a contractual, "merchant-sovereign" relationship [269] to a constitutional, "citizen-sovereign" relationship [204], where users move from customer to citizen, and platforms move from merchant to legitimizing and supporting a structure of democratic governance. This change in goals requires a shift in the perspectives of both users and platform operators.

7.4.2 A Normative Shift in User Roles

There are many aspects of the design of online discussion systems that push users towards the perspective of a customer dealing with a vendor. One is the decision by many major platforms to hide any moderation or curation that is going on behind the scenes. This includes large teams of paid moderators sifting through flagged posts oftentimes in locations far from where those posts were made for small amounts of money [281]. It also includes paid journalists or editors fact-checking content or crafting short blurbs for curated panels such as Twitter Moments or Facebook Trending News. To the typical user, it just magically happens that toxic content is not clogging up their feed or that some content has been distilled and fact-checked for them. This means that users themselves don't have to lift a finger to get these benefits. But when cracks form in the facade of effortless curation and users fall through, they realize that they also don't have power to change the status quo. For instance, when Facebook enforced a real-name policy on account names, ethnic groups such as Native Americans were incorrectly targeted and had little recourse to get their accounts reinstated.

While we can build end user curation systems, it still remains to be seen how often users would use those capabilities. For users accustomed to being catered to, this would involve a shift in perspective where they must acknowledge the invisible labor that goes into curation, appreciate its value, and commit to doing their part as a citizen of the space.

There are multiple ways to signal these user norms. The first, which this thesis addresses, is to **hand over greater control**, which demonstrates trust and respect towards users that they know how to self-govern. The second is to make the work of discussion curation more **visible to everyday users** instead of hiding it away [321]. This could help users understand the value that curation provides. For instance, in the Squadbox tool, we saw the need for relationship maintenance between owners and moderators since moderators weren't always sure they were doing a good job. A useful addition would be indicators or notices to owners of what their moderators had done for them so that owners could thank them for their work.

A third way is to actually **reward users** who do good curation work with greater social capital and editorial powers. Within Wikipedia, editors must be elected to curatorial positions such as administrator or bureaucrat after a public discussion within the community. While accessing these roles involves taking on greater responsibilities, they also signal a measure of respect from the community and an increase in social status. Similarly, within StackOverflow, users achieve the ability to edit other people's posts and other curation powers as they gain standing within the community. Given higher visibility, curators can also more effectively model good behavior to newcomers. There may even be cases where monetary rewards might support good curation work, such as paying discussion curators a portion of earnings on platforms such as Twitch or Youtube where content creators make money from advertisers or fans. This is an area that needs more research.

7.4.3 A Shift in Social Platform Thinking

In addition to users changing how they think about their role within online discussion systems, social platforms should also adopt different perspectives. A question that could be asked of online discussion systems is why haven't platform operators and developers already adopted more collective curation abilities in their systems?

Balance of Power Between Users and Platforms

One possibility is that platform operators have a fear of losing control over their user base and control over the ability to set the terms of what content users see. After all, when community members have greater collective power, they also gain leverage to push back against platform operators. This happened in the case of the Reddit "Blackout", where volunteer moderators of large subreddits on Reddit collectively turned their subreddits private to protest platform actions [225].

However, it is not necessarily a detrimental thing for platform operators if they lose some power relative to users. A more appropriate consideration would be what is the *right balance of responsibilities between different stakeholders* and *who is best tasked to perform what curation*. In my work, I argue that for most online discussion platforms, the current balance of power is tilted much too far away from users. This leads to problems as decision-makers become too removed from the people who are affected by their decisions.

One can see this negotiation play out for instance between the Wikimedia Foundation, the nonprofit organization supporting projects such as Wikipedia, and the Wikipedia community of editors. In 2013 the Wikimedia Foundation began development on a project called Flow [362], also known as Structured Discussions, that would fundamentally alter Wikipedia Talk Pages to look and feel more like typical forums, with structured comments and infinite scrolling, in order to make Talk Pages more accessible for newer editors. Reaction to the project was negative, particularly from more core editors in the Wikipedia community [94], who were used to the flexibility afforded by WikiText and had developed processes and tools on top of it. In the end, Flow was shelved on English Wikipedia in 2016 after a limited rollout. In this case, the failure of Flow can be traced back to not enough consultation from the people who would be most affected by Flow during its conception. The lesson, then, is for platforms to not unilaterally make decisions for their core users in cases where users are deeply affected. Instead, platforms could give users more tools to decide for themselves what they want.

The Value of Human Curation

A separate perspective is that platform operators have a fear of the human touch and the biases that could plausibly arise from it as opposed to automation, as well as a lack of belief in the value and consequence of curation itself.

This perspective may have explained the shifting decisions around the Facebook Trending News panel over the last few years. Facebook launched a "Trending" panel on the homepage sidebar in 2014, loosely based on a similar feature in Twitter. The panel featured handwritten headlines regarding each trending item and was manually curated by a team of journalists [49]. After concerns over bias, Facebook fired the editorial team in 2016 and replaced the manual curation with automation [366]. To be fair, much of this fear was stoked by outside forces, including conservative media, that accused Facebook of anti-conservative bias. And some of the concerns was warranted, given reports of the lack of diversity of the editorial team or transparency regarding their actions [247].

However, as mentioned earlier, automation comes with its own set of biases [22] and is not by default more or less biased than human curation. It is simply another form of editorial decision-making that has its own set of priorities. And in this case, without human oversight, the automated Trending panel began to recommend more misinformation, conspiracy theories, and highly partisan content [155]. In response to pressure from the government, media, and the public, Facebook scraped the feature. In more recent events, it appears that Facebook has come full circle and will once again be hiring a small team of journalists to curate a new "News Tab" [268].

Finally, Gilespie argues in his book "Custodians of the Internet" that "*platforms are not platforms without moderation*" [107]. That is, moderation is a central part of what makes a platform a platform, and a platform's moderation practices are a central component of what distinguishes it from another platform. Despite the consequence of curation decisions, many social platforms seem to regard them as an afterthought, hastily setting up contingencies or rewriting guidelines after the latest public relations emergency. Instead, systems for discussion should consider curation as equally valuable as content creation, and they should dedicate equal amounts of time to building tools, designing user workflows, and motivating contributions towards curation as they do towards content creation.

7.5 Conclusion

This discussion lays out several ways to put the contributions of the four systems in this thesis into a broader landscape of the design of curation tools. Implications arising from this research demonstrate ways that discussion curation artifacts should be designed to be maximally useful to discussion participants, re-visitors to the discussion, newcomers to the community, and other communities of practice. Discussion curation artifacts have value as a collaborative documentation of negotiation, a repository of organizational memory, a connector between different sociotechnical systems, and a translator between different communities.

Discussion curation artifacts have value but only if they accurately reflect the insights of end users. Unlike current common systems for discussion curation, we need systems that can reduce errors related to incorrect emulation of user preferences and that can improve the representation of end users in the curation process. One way to involve end users more is to incorporate more networked forms of curation that allows end users to communicate their needs but also gives them the ability to collaborate with others, reducing individual effort. More algorithmic forms of curation could also reduce individual effort but need to be developed so that end users working alone or together can effectively build and deploy such tools. Finally, curation decisions that require a group or platform scope can still empower end users by directly involving their voice in decision-making or allowing users to elect representatives.

To see this vision enacted, we also need to see normative change on the part of users and on the part of platform operators. Instead of hiding curation, it should be visible so that users are aware of what is being done on their behalf. Online discussion systems could also be designed to encourage and reward collective curation. Finally, platforms operators should realize that currently they hold too much power and also carry more responsibility than they can handle when it comes to conducting discussion curation for so many people. As builders of sociotechnical systems, we can and should re-imagine new online discussion platforms that value the needs of end users and that value human curation.

Chapter 8

Conclusion

This dissertation argues that the longstanding problems that online discussion systems face, including an overload of information and the presence of unwanted and harmful content, can be traced back to the failure of online discussion systems to innovate in the *ways that users can curate their discussions and discussion environments*. In response, this thesis presents four systems that demonstrate how end users could be provided with greater curation power, both individually and collectively.

8.1 Summary of Contributions

This thesis makes contributions towards how to design systems for collective discussion curation in a way that minimizes users' efforts and reflects users' needs. To that end, Chapter 2 traces the evolution of our current tools for online discussion and then synthesizes research from fields relating to collaborative and social computing, personal and collective information management, and information visualization and automatic processing to inform the design of novel discussion curation tools.

Wikum (Chapter 3) explores how online forums could tackle scale through *collaborative summarization* and presents a novel recursive summarization process for users to build on each other's work as well as a novel summary tree artifact for exploring discussions at different levels of summarization. I present a case study of editors on Wikipedia who already perform the work of summarizing discourse but have difficulty doing so without effective tools or the ability to collaborate. From a deployment of Wikum with these editors, I find evidence suggesting the tool's usefulness for distributing cognitive load across time and people.

In Chapter 4, I examine group chat systems, finding users overburdened with trying to catch up and wanting ways to get structured, contextual updates. I develop lightweight techniques for *teamsourced notetaking and tagging* within chat and integrate them into *Tilda*, a tool for marking up chat in situ. Deployments of Tilda with Slack teams demonstrated active use of the tool for both marking up chat and catching up using chat summaries.

Moving from curation that alters discussion, I turn towards systems for curating message *workflows*. In Chapter 5, I study communities of mailing lists, finding tensions between members due to mismatches in posting behaviors and desires. I develop a new mailing list system called *Murmur* focused on distributed *fine-grained delivery customization* by senders and receivers.

In Chapter 6, I examine the problem of online harassment, finding users overwhelmed and turning to friends for help who can understand their contextual needs. In response, I develop *Squadbox*, a system for *friendsourced moderation* of email, where people faced with a harassment campaign can turn on Squadbox and redirect potentially harassing messages to their friends who can vet the messages for them.

Finally, Chapter 7 presents a series of implications for the design of discussion curation artifacts and places the presented systems into broader frameworks examining scope, automation, and representation to characterize discussion curation systems as a whole.

The contributions fall into the following three main research areas of social computing, computer-supported cooperative work (CSCW), and crowdsourcing.

- **Social Computing**: Distributed social arrangements such as friendsourced moderation and teamsourced curation permit new forms of collaborative discussion curation across different social relationships.

- **CSCW**: As a form of collaborative work, these discussion curation tools offer new ways of doing, such as superimposing structure, and new artifact designs, such as summary trees, for negotiating boundaries and routinizing work, towards improving collaboration.

- **Crowdsourcing**: Novel workflows such as recursive summarization effectively distribute and combine inputs for collective curation. Distributed processes such as fine-grained delivery customization explore collective intelligence towards information management.

8.2 Future Work

This thesis presents four examples of collective discussion curation tools but there are many more that could be explored in future research. There are also additional dimensions than the ones examined in this thesis that could be considered. In this section, I describe some of the limitations of this thesis and promising areas for future work.

8.2.1 Designing for User Roles, Life Cycles, and Incentives

In my thesis, I did not focus on designing user roles and incentives. In each system, each user of the system has the ability to perform any curation action that was available. There are a few exceptions, namely owners of Wikum projects can block individuals or give edit versus read privileges, and Squadbox owners can determine what their moderators can do for them. But other than that, there was no concept of particular user roles or hierarchies of power. Instead, systems could be designed to support different explicit roles or support the formation of emergent roles.

In addition, I did not explore the possibility of transitions over a user's life cycle. For instance, the needs of a newcomer to a system are different from the needs of seasoned community member. Users usually don't start out with the most complicated or consequential curation work but work their way up [13]. Systems could be designed to support the seamless transition of users as they gain social capital and take on greater curatorial responsibilities. In this work, I did not pursue any long term field studies. Such studies would be useful to understand how users of a discussion curation system alter their behavior over longer periods of time.

Finally, the transition of users from casual to central contributors depends on systems that can sufficiently motivate and incentivize participation. While this work examined motivations in the deployments of tools to real users, more work is necessary to understand how to design different incentives for curation. Potential avenues that have been explored in the past in other areas include personal or social gamification, just-in-time nudges, social proof, and financial incentives.

8.2.2 Governance Models

An important thread throughout this work has been on curation's relationship to governance and how social platforms can gain legitimacy for their curatorial decisions. Today, large social platforms grappling with problems of speech have build up large structures of governance strikingly reminiscent of existing institutions offline [182]. Most recently, Mark Zuckerberg has called for a Facebook "Supreme Court" that would weigh in on difficult content moderation cases [387]. There is a great deal of existing research on governance in the offline setting that may be a useful way to consider platform governance. For instance, I am collaborating on a project examining constitutional jury systems such as exist in the U.S. and applying those models to online content moderation. Future research could explore how other offline governance processes might translate to the online setting. Systems could be developed that enable democratic governance of code, whether that be code for interfaces or code for algorithms.

8.2.3 Improving Discussant Experiences While Conversing

While Wikum supports understanding of discourse after the fact, and Murmur and Squadbox support moderating out unwanted discourse, I did not address how to design systems to improve the quality of discourse for *discussants during discussion.*

Tilda is the only tool that can be used while discussions are ongoing. We did not explicitly study the effect of using Tilda on discussions; however, one interesting post-study comment from a Tilda user mentioned how they noticed discussions improving after the addition of Tilda in that they drove towards a goal faster. The presence of the Tilda discourse act categories signaled to users to provide statements conforming to those categories. This suggestion warrants further study. More broadly, it would be interesting to explore how discussion systems can give participants a sense of progress, even when there is no tangible output. One way could be to treat the discussion artifact itself as an output or an input into another artifact, as opposed to something discarded once over.

Another important and related area of improvement is how we can increase empathy across sides through discourse, particularly given the prevalence of polarized discourse readily observed online today. I hypothesize that an important aspect of breaking through echo chambers is *framing*, particularly the underlying moral frames behind arguments. Early experiments I have collaborated on in this area provided evidence that getting discussants to share their moral values with others increased empathy towards the other side [350]. Work I have done to externalize, detect [379, 380], and visualize [67] moral values could give people a better frame for interpreting different opinions. Another possibility is that collaborative discussion summarization itself could also be a mechanism for encouraging empathy. Currently, Wikum supports only discussion summarization after a discussion is over. An open question is how to design discussion systems that embed summarization into the process of discussion.

8.2.4 Building Beyond Textual Discourse

This thesis only explored textual discourse online. But today, people converse using a variety of media, include streaming video, images, GIFs and other memes, emojis, and more. Some of these media, such as livestream videos, have even greater issues with overload as they are harder to skim. There is also spoken discourse, both offline and online, using phone, video, or in-person meetings between small groups. Some techniques for notetaking and tagging that were explored in Tilda could be applied towards synthesizing spoken communication, given improvements in speech-to-text technology. Finally, future research could consider software to support larger gatherings such as conferences online, where many discussions might be going on simultaneously, and how to curate those discussions.

8.2.5 Scaling Up Discussion Synthesis

The techniques introduced in Wikum and Tilda are just two of many ways that people can provide signals to synthesize and add structure to discussion. As an example of another signal, thousands of readers may explore the same forum, creating paths of interest through different threads, yet each newcomer must start again from scratch. Harnessing both active and passive signals will be helpful towards building web-scale systems for synthesis and exploration of discussion. For instance, what is the best way to build the Wikipedia of opinions or the Google of public forums?

Another way to consider scaling up is to develop a set of common patterns of discussion structures. The discourse actions that people take while conversing can vary greatly even though discussion threads have little variation in appearance today. For instance, the existence of structured Q&A sites suggests that Q&A discussions should have a different presentation than other types of discussions such as pro-con deliberations. Building on work I conducted on automatically characterizing common discourse act chains [382], future research could explore new representations of discourse beyond simply threaded and non-threaded, towards a broad typology of discourse structures.

8.2.6 From Users Curating to Users Creating and Remixing Curation Tools

In this thesis, I conducted needfinding studies with potential users before developing systems to address their needs using a user-centered approach. However, in each of the current systems, users can use the features that I developed but have little ability to develop other forms of curation on their own or affect how the interfaces look and behave. In addition, I define a particular data schema and while users are able to decide where and to what depth they wish to apply that schema, they cannot alter the schema itself. For instance, users can in Murmur perform a series of pre-defined customizations as senders or receivers, such as adding or blocking participants or tags. However, they are not free to define their own customizations, for example to block a person only on weekdays.

An interesting future line of research would be to consider software systems and collaborative workflows that could better facilitate *participatory design* or *co-design* at scale for online discussion systems [290]. Another line of work would be to develop systems where users can create their own data schemas and author their own customizations on top of those schemas. From there, systems could allow users to share their customizations with other users, who could then remix multiple people's work to create something new. The research question of how to design such a system so that it is flexible yet accessible to end users is one that we are currently exploring in the realm of email customization [260].

When we see “internet of things”, let’s make it an internet of beings.
When we see “virtual reality”, let’s make it a shared reality.
When we see “machine learning”, let’s make it collaborative learning.
When we see “user experience”, let’s make it about human experience.
When we hear “the singularity is near”, let us remember: the Plurality is here.

–Audrey Tang

Bibliography

[1] Mark S Ackerman. The intellectual challenge of cscw: the gap between social requirements and technical feasibility. *Human–Computer Interaction*, 15(2-3):179–203, 2000.

[2] Mark S Ackerman, Juri Dachtera, Volkmar Pipek, and Volker Wulf. Sharing knowledge and expertise: The CSCW view of knowledge management. *Computer Supported Cooperative Work (CSCW)*, 22(4-6):531–573, 2013.

[3] Mark S Ackerman and Christine Halverson. Organizational memory as objects, processes, and trajectories: An examination of organizational memory in use. *Computer Supported Cooperative Work (CSCW)*, 13(2):155–189, 2004.

[4] Mark S Ackerman and Thomas W Malone. *Answer Garden: A tool for growing organizational memory*, volume 11. ACM, 1990.

[5] Mark S Ackerman, Anne Swenson, Stephen Cotterill, and Kurtis DeMaagd. I-diag: from community discussion to knowledge distillation. In *Communities and Technologies*, pages 307–325. Springer, 2003.

[6] CJ Adams and Lucas Dixon. Better discussions with imperfect models. `https://medium.com/the-false-positive/better-discussions-with-imperfect-models-91558235d442` [Last accessed: January 3, 2018], September 2017.

[7] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *WSDM '08*, pages 183–194. ACM, 2008.

[8] June Ahn, Brian S Butler, Cindy Weng, and Sarah Webster. Learning to be a better Q'er in social Q&A sites: social norms and information artifacts. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*, page 4. American Society for Information Science, 2013.

[9] Tarfah Alrashed, Ahmed Hassan Awadallah, and Susan Dumais. The lifetime of email messages: A large-scale analysis of email revisitation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 120–129, New York, NY, USA, 2018. ACM.

[10] Paul André, Aniket Kittur, and Steven P Dow. Crowd synthesis: Extracting categories and clusters from complex data. In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing*, pages 989–998. ACM, 2014.

[11] Paul André, Robert E Kraut, and Aniket Kittur. Effects of simultaneous and sequential work structures on distributed collaborative interdependent tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 139–148. ACM, 2014.

[12] Pablo Aragón, Andreas Kaltenbrunner, Antonio Calleja-López, Andrés Pereira, Arnau Monterde, Xabier E Barandiaran, and Vicenç Gómez. Deliberative platform design: The case study of the online discussions in decidim barcelona. In *International Conference on Social Informatics*, pages 277–287. Springer, 2017.

[13] Ofer Arazy, Felipe Ortega, Oded Nov, Lisa Yeo, and Adam Balila. Functional roles and career paths in Wikipedia. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1092–1105. ACM, 2015.

[14] Zahra Ashktorab and Jessica Vitak. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 3895–3905, New York, NY, USA, 2016. ACM.

[15] John Langshaw Austin. *How to do things with words.* Oxford University Press, 1975.

[16] John Perry Barlow. Declaration of independence for cyberspace, 1996.

[17] Victoria Bellotti, Nicolas Ducheneaut, Mark Howard, and Ian Smith. Taking email to task: the design and evaluation of a task management centered email tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 345–352, New York, NY, USA, 2003. ACM.

[18] Yochai Benkler. *The wealth of networks: How social production transforms markets and freedom.* Yale University Press, 2006.

[19] Frank Bentley, Nediyana Daskalova, and Nazanin Andalibi. If a person is emailing you, it just doesn't make sense: Exploring changing consumer behaviors in email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '17, pages 85–95, New York, NY, USA, 2017. ACM.

[20] Michael Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. Quantifying the invisible audience in social networks. *Proc. CHI*, pages 21–30, 2013.

[21] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. Soylent: a

word processor with a crowd inside. *Communications of the ACM*, 58(8):85–94, 2015.

[22] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II*, pages 405–415. Springer, Springer International Publishing, 2017.

[23] Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, and Anand Swaminathan. Mining email social networks. In *Proceedings of the 2006 international workshop on Mining software repositories*, MSR '06, pages 137–143, New York, NY, USA, 2006. ACM.

[24] Pernille Bjørn. New fundamentals for CSCW research: from distance to politics. *interactions*, 23(3):50–53, 2016.

[25] Laura W Black, Howard T Welser, Dan Cosley, and Jocelyn M DeGroot. Self-governance through group discussion in Wikipedia: Measuring deliberation in online groups. *Small Group Research*, 42(5):595–634, 2011.

[26] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and its consequences for online harassment: Design insights from heartmob. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):24:1–24:19, December 2017.

[27] Anton Bogdanovych, Helmut Berger, Simeon Simoff, and Carles Sierra. Narrowing the gap between humans and agents in e-commerce: 3d electronic institutions. In *EC-Web*, volume 5, pages 128–137. Springer, 2005.

[28] Michael L. Bourke and Sarah W. Craun. Secondary traumatic stress among internet crimes against children task force personnel: Impact, risk factors, and coping strategies. *Sexual Abuse*, 26(6):586–609, 2014.

[29] Danah M Boyd and Nicole B Ellison. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13(1):210–230, 2007.

[30] David Brumley. Tracking hackers on IRC. *Usenix*, November 1999. Available: `https://www.usenix.org/legacy/publications/login/1999-11/features/hackers.html` [Last accessed: 2017-09-14].

[31] Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. Towards an ISO standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2010.

[32] Moira Burke and Robert Kraut. Mopping up: modeling Wikipedia promotion decisions. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 27–36. ACM, 2008.

[33] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in Wikipedia. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1101–1110. ACM, 2008.

[34] Carrie J Cai, Shamsi T Iqbal, and Jaime Teevan. Chain reactions: The impact of order on microtask chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3143–3154. ACM, 2016.

[35] Ann Frances Cameron and Jane Webster. Unintended consequences of emerging communication technologies: Instant messaging in the workplace. *Computers in Human behavior*, 21(1):85–103, 2005.

[36] Joyce Chai, Veronika Horvath, Nicolas Nicolov, Margo Stys, Nanda Kambhatla, Wlodek Zadrozny, and Prem Melville. Natural language assistant: A dialog system for online product recommendation. *AI Magazine*, 23(2):63, 2002.

[37] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3175–3187, New York, NY, USA, 2017. ACM.

[38] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis*. Sage, 2006.

[39] Kathy Charmaz and Linda Liska Belgrave. Grounded theory. *The Blackwell encyclopedia of sociology*, 2007.

[40] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How community feedback shapes user behavior. 2014.

[41] Ed H Chi, Peter Pirolli, Kim Chen, and James Pitkow. Using information scent to model user information needs and actions and the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 490–497. ACM, 2001.

[42] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1999–2008. ACM, 2013.

[43] Patrick Chiu, John Boreczky, Andreas Girgensohn, and Don Kimber. Liteminutes: an internet-based system for multimedia meeting minutes. In

Proceedings of the 10th international conference on World Wide Web, pages 140–149. ACM, 2001.

[44] Patrick Chiu, Ashutosh Kapuskar, Sarah Reitmeier, and Lynn Wilcox. Notelook: Taking notes in meetings with digital video and ink. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 149–158. ACM, 1999.

[45] Mercia Coetzee, Annette Wilkinson, and Daleen Krige. Mapping the social media landscape: a profile of tools, applications and key features. 2016.

[46] William W Cohen, Vitor R Carvalho, and Tom M Mitchell. Learning to classify email into "speech acts". In *EMNLP '04*, pages 309–316, 2004.

[47] Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. Finding question-answer pairs from online forums. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 467–474. ACM, 2008.

[48] Josh Constine. Facebook messenger hits 1.2 billion monthly users, up from 1b in july. *TechCrunch*, 12 April 2017. Available: `https://techcrunch.com/2017/04/12/messenger/` [Last accessed: 2017-09-14].

[49] Josh Constine. Facebook launches trending topics on web with descriptions of why each is popular, Jan 2014.

[50] Susan Corbett. Curation nation: how to win in a world where consumers are creators, 2011.

[51] Juliet Corbin and Anselm Strauss. *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Sage, 2008.

[52] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. Calendar.help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2382–2393. ACM, 2017.

[53] Justin Cranshaw and Aniket Kittur. The polymath project: lessons from a successful online collaboration in mathematics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1865–1874. ACM, 2011.

[54] Kate Crawford and Tarleton Gillespie. What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society*, 2014.

[55] Laurie Cubbison. Configuring listserv, configuring discourse. *Computers and Composition*, 16(3):371–381, 1999.

[56] Mary Czerwinski, Edward Cutrell, and Eric Horvitz. Instant messaging and interruption: Influence of task type on performance. In *OZCHI 2000 conference proceedings*, volume 356, pages 361–367, 2000.

[57] Laura Dabbish and Robert Kraut. Email overload at work. *Proc. CSCW*, pages 431–440, 2006.

[58] Laura A. Dabbish, Robert E. Kraut, Susan Fussell, and Sara Kiesler. Understanding email use: predicting action on a message. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 691–700, New York, NY, USA, 2005. ACM.

[59] Richard L Daft and Robert H Lengel. Organizational information requirements, media richness and structural design. *Management science*, 32(5):554–571, 1986.

[60] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708. ACM, 2012.

[61] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. *Association for Computational Linguistics*, 2013.

[62] Srayan Datta and Eytan Adar. Extracting inter-community conflicts in reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 146–157, 2019.

[63] Kushal Dave, Martin Wattenberg, and Michael Muller. Flash forums and forumreader: navigating a new kind of large-scale online discussion. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, pages 232–241. ACM, 2004.

[64] Richard C Davis, James A Landay, Victor Chen, Jonathan Huang, Rebecca B Lee, Frances C Li, James Lin, Charles B Morrey III, Ben Schleimer, Morgan N Price, et al. Notepals: Lightweight note sharing by the group, for the group. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 338–345. ACM, 1999.

[65] Debian. Meetbot. *Debian*, 8 Janary 2017. Available: `https://wiki.debian.org/MeetBot` [Last accessed: 2017-09-08].

[66] Utpal M Dholakia, Richard P Bagozzi, and Lisa Klein Pearo. A social influence model of consumer participation in network-and small-group-based virtual communities. *International journal of research in marketing*, 21(3):241–263, 2004.

[67] Nicholas Diakopoulos, Amy X Zhang, Dag Elgesem, and Andrew Salway. Identifying and analyzing moral evaluation frames in climate change blog discourse. In *ICWSM*, 2014.

[68] Digest.AI. `https://slackdigest.com`.

[69] Jill P. Dimond, Michaelanne Dye, Daphne Larose, and Amy S. Bruckman. Hollaback!: The role of storytelling online in a social movement organization. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 477–490, New York, NY, USA, 2013. ACM.

[70] Darcy DiNucci. Fragmented future. *Print*, 53(4):32–33, 1999.

[71] Discourse. `https://www.discourse.org`.

[72] Disqus. `https://disqus.com`.

[73] Judith Donath, Karrie Karahalios, and Fernanda Viegas. Visualizing conversation. *Journal of Computer-Mediated Communication*, 4(4):0–0, 1999.

[74] Robert T Douglass, Mike Little, and Jared W Smith. *Building online communities with Drupal, phpBB, and WordPress*. Springer, 2006.

[75] Mark Dredze, Tessa Lau, and Nicholas Kushmerick. Automatically classifying emails into activities. In *Proceedings of the 11th international conference on Intelligent user interfaces*, IUI '06, pages 70–77, New York, NY, USA, 2006. ACM.

[76] Mark Dredze, Bill N. Schilit, and Peter Norvig. Suggesting email view filters for triage and search. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence*, IJCAI '09, pages 1414–1419, Palo Alto, CA, USA, 2009. AAAI.

[77] Maeve Duggan. Online harassment 2017. *The Pew Research Center*. Available: `http://www.pewinternet.org/2017/07/11/online-harassment-2017/` [Last accessed: September 8, 2017], July 2017.

[78] Nicole Ellison, Charles Steinfield, and Cliff Lampe. The benefits of Facebook “friends". *JCMC*, 12(4):1143–1168, 2007.

[79] Nicole B Ellison and Danah M Boyd. Sociality through social network sites. In *The Oxford handbook of internet studies*. 2013.

[80] Steven Englehardt, Jeffrey Han, and Arvind Narayanan. I never signed up for this! privacy implications of email tracking. *Proceedings on Privacy Enhancing Technologies*, 1:109–126, 2018.

[81] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, pages 457–479, 2004.

[82] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. I always assumed that i wasn't really that close to [her]: Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 153–162. ACM, 2015.

[83] Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, and Ken Goldberg. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1175–1184. ACM, 2010.

[84] Robert Faris, Amar Ashar, Urs Gasser, and Daisy Joo. Understanding harmful speech online. *Berkman Klein Center Research Publication 2016-21*, December 2016.

[85] Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S Bernstein. Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 473. ACM, 2018.

[86] Daniel C Feldman. The development and enforcement of group norms. *Academy of management review*, 9(1):47–53, 1984.

[87] Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. Learning to detect conversation focus of threaded discussions. In *NAACL HLT '06*, pages 208–215. ACL, 2006.

[88] Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. Behind the article: Recognizing dialog acts in Wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786. Association for Computational Linguistics, 2012.

[89] Sara Fischer and Alison Snyder. Exclusive poll: America sours on social media giants. *Axios.* Available: `https://www.axios.com/america-sours-on-social-media-giants-1542234046-c48fb55b-48d6-4c96-9ea9-a36e80ab5deb.html` [Last accessed: November 19, 2018], November 2018.

[90] Danyel Fisher, A. Brush, Eric Gleave, and Marc Smith. Revisiting Whittaker & Sidner's "email overload" ten years later. *Proc. CSCW*, pages 309–312, 2006.

[91] Andrew J Flanagin and Miriam J Metzger. Internet use in the contemporary media environment. *Human communication research*, 27(1):153–181, 2001.

[92] Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.

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

[94] Wikimedia Foundation. Collaboration/flow satisfaction survey/report. https://meta.wikimedia.org/wiki/Collaboration/Flow_satisfaction_survey/Report.

[95] Jesse Fox and Wai Yen Tang. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media & Society*, 19(8):1290–1307, 2017.

[96] Siwei Fu, Jian Zhao, Hao Fei Cheng, Haiyi Zhu, and Jennifer Marlow. T-cal: Understanding team conversational data with calendar-based visualization. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 500. ACM, 2018.

[97] Daniel Funke. Wikipedia vandalism could thwart hoax-busting on Google, YouTube and Facebook. *Poynter*. Available: https://www.poynter.org/fact-checking/2018/wikipedia-vandalism-could-thwart-hoax-busting-on-google-youtube-and-facebook/ [Last accessed: August 10, 2019], 2018.

[98] Francis Galton. Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451, 1907.

[99] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90. Association for Computational Linguistics, 2017.

[100] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 340–348. Association for Computational Linguistics, 2010.

[101] R Kelly Garrett and James N Danziger. IM = interruption management? instant messaging and disruption in the workplace. *Journal of Computer-Mediated Communication*, 13(1):23–42, 2007.

[102] R Stuart Geiger. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society*, 19(6):787–803, March 2016.

[103] R Stuart Geiger and Heather Ford. Participation in Wikipedia's article deletion processes. In *Proceedings of the 7th international symposium on Wikis and open collaboration*, pages 201–202. ACM, 2011.

[104] R Stuart Geiger, Nelle Varoquaux, Charlotte Mazel-Cabasse, and Chris Holdgraf. The types, roles, and practices of documentation in data analytics open source software libraries. *Computer Supported Cooperative Work (CSCW)*, 27(3-6):767–802, 2018.

[105] Werner Geyer, Heather Richter, and Gregory D Abowd. Towards a smarter meeting record-capture and access of meetings revisited. *Multimedia Tools and Applications*, 27(3):393–410, 2005.

[106] Eric Gilbert. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 803–808. ACM, 2013.

[107] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press, 2018.

[108] Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, pages 229–233, New York, NY, USA, 2017. ACM.

[109] Google. Google groups. `https://groups.google.com`.

[110] Google. Google hangouts chat. `https://chat.google.com`.

[111] Google. Perspective API. *Google Jigsaw.* Available: `https://www.perspectiveapi.com/` [Last accessed: September 8, 2017], 2017.

[112] Sallie Gordon, Jill Gustavel, Jana Moore, and Jon Hankey. The effects of hypertext on reader knowledge representation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 32, pages 296–300. SAGE Publications, 1988.

[113] Timothy Gowers and Michael Nielsen. Massively collaborative mathematics. *Nature*, 461(7266):879–881, 2009.

[114] Sukeshini A. Grandhi and Lyndsey K. Lanagan-Leitzel. To reply or to reply all: Understanding replying behavior in group email communication. In *Proceedings of the 2016 ACM conference on Computer-supported cooperative work.*, CSCW '16, pages 560–569, New York, NY, USA, 2016. ACM.

[115] Antonietta Grasso and Gregorio Convertino. Collective intelligence in organizations: Tools and studies. *Computer Supported Cooperative Work (CSCW)*, 21(4-5):357–369, 2012.

[116] Alan Gray and Mads Haahr. Personalised, collaborative spam filtering. In *CEAS*, 2004.

[117] Irene Greif. Computer-supported cooperative work: A book of readings. 1988.

[118] Catherine Grevet, David Choi, Debra Kumar, and Eric Gilbert. Overload is overloaded: email in the age of gmail. CHI '14, pages 793–802. ACM, 2014.

[119] James Grimmelmann. The virtues of moderation. *Yale JL & Tech.*, 17:42, 2015.

[120] Jonathan Grudin. Computer-supported cooperative work: History and focus. *Computer*, 27(5):19–26, 1994.

[121] Jonathan Grudin. Groupware and social dynamics: Eight challenges for developers. In *Readings in Human–Computer Interaction*, pages 762–774. Elsevier, 1995.

[122] Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. The knowledge accelerator: Big picture thinking in small pieces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2258–2270. ACM, 2016.

[123] Nathan Hahn, Joseph Chee Chang, and Aniket Kittur. Bento browser: Complex mobile search without tabs. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 251. ACM, 2018.

[124] Aaron Halfaker and John Riedl. Bots and cyborgs: Wikipedia's immune system. *Computer*, 45(3):79–82, 2012.

[125] Hazel Hall and Dianne Graham. Creation and recreation: motivating collaboration to generate knowledge capital in online communities. *International Journal of Information Management*, 24(3):235–246, 2004.

[126] Mark Handel and James D Herbsleb. What is chat doing in the workplace? In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 1–10. ACM, 2002.

[127] Derek L Hansen, Mark S Ackerman, Paul J Resnick, and Sean Munson. Virtual community maintenance with a collaborative repository. *Proceedings of the American Society for Information Science and Technology*, 44(1):1–20, 2007.

[128] Garrett Hardin. The tragedy of the commons. *science*, 162(3859):1243–1248, 1968.

[129] Randi Lee Harper. Good game auto blocker. Available: `https://github.com/freebsdgirl/ggautoblocker` [Last accessed: 2017-09-08], 2014.

[130] Randi Lee Harper. Putting out the Twitter trashfire. Art + Marketing, `https://artplusmarketing.com/putting-out-the-twitter-trashfire-3ac6cb1af3e` [Last accessed: September 8, 2017], February 2016.

[131] Sandra Harrison. E-mail discussions as conversation: moves and acts in a sample from a listserv discussion. *Linguistik online*, 1(4), 1998.

[132] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[133] James Hartley. Note-taking research: Resetting the scoreboard. *Bulletin of the British Psychological Society*, 1983.

[134] Del Harvey and David Gasca. Serving healthy conversations, May 2018.

[135] Kazi Saidul Hasan and Vincent Ng. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *EMNLP '14*, pages 751–762, 2014.

[136] James D Herbsleb, David L Atkins, David G Boyer, Mark Handel, and Thomas A Finholt. Introducing instant messaging and chat in the workplace. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 171–178. ACM, 2002.

[137] William C Hill, James D Hollan, Dave Wroblewski, and Tim McCandless. Edit wear and read wear. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 3–9. ACM, 1992.

[138] William Fawcett Hill. Learning through discussion. Beverly Hills. *CA: Sage. Hunt, P.(1951). The case method of instruction. Harvard Educational Review*, 21:2–19, 1969.

[139] Starr R Hiltz and Murray Turoff. Structuring computer-mediated communication systems to avoid information overload. *Communications of the ACM*, 28(7):680–689, 1985.

[140] Marit Hinnosaar, Toomas Hinnosaar, Michael Kummer, and Olga Slivko. Wikipedia matters: a significant impact of user-generated content on real-life choices, 2017.

[141] HipChat. `https://www.hipchat.com`.

[142] Arlie Russell Hochschild. Emotion work, feeling rules, and social structure. *American journal of sociology*, 85(3):551–575, 1979.

[143] Jacob Hoffman-Andrews. Blocktogether. Available: `https://blocktogether.org/` [Last accessed: 2017-09-08], 2017.

[144] Liangjie Hong and Brian D Davison. A classification-based approach to question answering in discussion boards. In *SIGIR '09*, pages 171–178. ACM, 2009.

[145] Enamul Hoque and Giuseppe Carenini. Convis: A visual text analytic system for exploring blog conversations. In *Computer Graphics Forum*, volume 33, pages 221–230. Wiley Online Library, 2014.

[146] Enamul Hoque and Giuseppe Carenini. Convisit: Interactive topic modeling for exploring asynchronous online conversations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 169–180. ACM, 2015.

[147] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving Google's perspective API built for detecting toxic comments. *CoRR*, abs/1702.08138, 2017.

[148] Yifeng Hu, Jacqueline Fowler Wood, Vivian Smith, and Nalova Westbrook. Friendships through im: Examining the relationship between instant messaging and intimacy. *Journal of Computer-Mediated Communication*, 10(1):00–00, 2004.

[149] Samuel Hulick. I used to be obsessed with slack but now i'm dropping it completely - here's why. *Business Insider*, 1 March 2016. Available: `http://www.businessinsider.com/i-used-to-be-obsessed-with-slack-but-now-im-dropping-it-completely-heres-why-2016-3` [Last accessed: 2017-09-09].

[150] Edward Hung. Deduction of Procmail Recipes from Classified Emails. *CMSC724 Database Management Systems, individual research project report*, 2001.

[151] Edwin Hutchins. *Cognition in the Wild.* Number 1995. MIT press, 1995.

[152] Avi Hyman. Twenty years of listserv as an academic tool. *The Internet and higher education*, 6(1):17–24, 2003.

[153] IFTTT. IFTTT Gmail, 2010. `https://ifttt.com/gmail`.

[154] Jane Im, Amy X Zhang, Christopher J Schilling, and David Karger. Deliberation and resolution on Wikipedia: A case study of requests for comments. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):74, 2018.

[155] Sasha Ingber. Facebook is scrapping its troubled 'trending' news section, June 2018.

[156] Shamsi T Iqbal and Eric Horvitz. Disruption and recovery of computing tasks: field study, analysis, and directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 677–686. ACM, 2007.

[157] Samuel P. Jacobs. How e-mail killer Slack will change the future of work. *Time*, 28 October 2015. Available: `http://time.com/4092354/how-e-mail-killer-slack-will-change-the-future-of-work/` [Last accessed: 2017-09-08].

[158] Sirkka L Jarvenpaa and Dorothy E Leidner. Communication and trust in global virtual teams. *Journal of Computer-Mediated Communication*, 3(4):JCMC346, 1998.

[159] Adrianne Jeffries. We're taking a break from Slack. Here's why. *Motherboard*, 16 May 2016. Available: https://motherboard.vice.com/en_us/article/aekk85/were-taking-a-break-from-slack-heres-why [Last accessed: 2017-09-09].

[160] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):12, 2018.

[161] Robert Johansen. *Groupware: Computer support for business teams.* The Free Press, 1988.

[162] William Jones. The future of personal information management, part I: our information, always and forever. *Synthesis lectures on information concepts, retrieval, and services*, 4(1):1–125, 2012.

[163] Sanjay Kairam, Mike Brzozowski, David Huffaker, and Ed Chi. Talking in circles. *Proc. CHI*, pages 1065–1074, 2012.

[164] Matthew Kam, Jingtao Wang, Alastair Iles, Eric Tse, Jane Chiu, Daniel Glaser, Orna Tarshish, and John Canny. Livenotes: a system for cooperative and augmented note-taking in lectures. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 531–540. ACM, 2005.

[165] Steven Karau and Kipling Williams. Social loafing. *Journal of personality and social psychology*, 65(4):681, 1993.

[166] Shih-Wen Ke, Chris Bowerman, and Michael Oakes. Perc: A personal email classifier. *European Conference on Information Retrieval*, pages 460–463, 2006.

[167] George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. Technology solutions to combat online harassment. In *Proceedings of the First Workshop on Abusive Language Online*, pages 73–77. Association for Computational Linguistics, 2017.

[168] Andruid Kerne, Nic Lupfer, Rhema Linder, Yin Qu, Alyssa Valdez, Ajit Jain, Kade Keith, Matthew Carrasco, Jorge Vanegas, and Andrew Billingsley. Strategies of free-form web curation: Processes of creative engagement with prior work. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, pages 380–392. ACM, 2017.

[169] Bernard Kerr. Thread arcs: An email thread visualization. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pages 211–218. IEEE, 2003.

[170] Mary Beth Kery, Marissa Radensky, Mahima Arya, Bonnie E John, and Brad A Myers. The story in the notebook: Exploratory data science using a literate programming tool. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 174. ACM, 2018.

[171] Kialo. `https://kialo.com`.

[172] Marjorie D Kibby. Email forwardables: folklore in the age of the internet. *New Media & Society*, 7(6):770–790, 2005.

[173] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge, MA*, 2012.

[174] Kenneth A Kiewra. Investigating notetaking and review: A depth of processing alternative. *Educational Psychologist*, 20(1):23–32, 1985.

[175] Su Nam Kim, Li Wang, and Timothy Baldwin. Tagging and linking web forum posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 192–202. Association for Computational Linguistics, 2010.

[176] Margaret R Kirkland and Mary Anne P Saunders. Maximizing student performance in summary writing: Managing cognitive load. *Tesol Quarterly*, 25(1):105–121, 1991.

[177] Aniket Kittur and Robert E Kraut. Beyond Wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 215–224. ACM, 2010.

[178] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1301–1318, New York, NY, USA, 2013. ACM.

[179] Aniket Kittur, Bongwon Suh, and Ed H Chi. Can you ever trust a wiki?: impacting perceived trustworthiness in Wikipedia. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, pages 477–480. ACM, 2008.

[180] Aniket Kittur, Bongwon Suh, Bryan A Pendleton, and Ed H Chi. He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462. ACM, 2007.

[181] Mark Klein. How to harvest collective wisdom on complex problems: An introduction to the mit deliberatorium. *Center for Collective Intelligence working paper*, 2011.

[182] Kate Klonick. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.*, 131:1598, 2017.

[183] Nicolas Kokkalis, Chengdiao Fan, Johannes Roith, Michael S. Bernstein, and Scott Klemmer. Myriadhub: Efficiently scaling personalized email conversations with valet crowdsourcing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 73–84, New York, NY, USA, 2017. ACM.

[184] Nicolas Kokkalis, Thomas Köhn, Carl Pfeiffer, Dima Chornyi, Michael S. Bernstein, and Scott R. Klemmer. Emailvalet: Managing email overload through private, accountable crowdsourcing. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1291–1300, New York, NY, USA, 2013. ACM.

[185] Joseph Konstan, Bradley Miller, David Maltz, Jonathan Herlocker, Lee Gordon, and John Riedl. Grouplens. *Communications of the ACM*, 40(3):77–87, 1997.

[186] Steve Kovach. I figured out a way to kill work email once and for all. *Venture Beat*, 19 February 2015. Available: `https://venturebeat.com/2016/10/20/slack-passes-4-million-daily-users-and-1-25-million-paying-users/` [Last accessed: 2017-09-08].

[187] Travis Kriplean, Ivan Beschastnikh, David W McDonald, and Scott A Golder. Community, consensus, coercion, control: cs*w or how policy mediates mass participation. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 167–176. ACM, 2007.

[188] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 265–274. ACM, 2012.

[189] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Andrew Ko. Is this what you meant?: promoting listening on the web with reflect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1559–1568. ACM, 2012.

[190] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 933–943. International World Wide Web Conferences Steering Committee, 2018.

[191] Greg Lambert. Where do listservs fit in a social media world? *American Association of Law Libraries Spectrum*, 13:8–13, 2009.

[192] Cliff Lampe and Paul 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. ACM, 2004.

[193] Cliff Lampe, Rick Wash, Alcides Velasquez, and Elif Ozkaya. Motivations to participate in online communities. *Proc. CHI*, pages 1927–1936, 2010.

[194] James A. Landay and Richard C. Davis. Making sharing pervasive: Ubiquitous computing for shared note taking. *IBM Systems Journal*, 38(4):531–550, 1999.

[195] David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. In *ICWSM*, pages 177–184, 2011.

[196] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 23–34. ACM, 2012.

[197] Walter S. Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P. Bigham. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, pages 151–162. ACM, 2013.

[198] Christoph Lattemann and Stefan Stieglitz. Framework for governance in open source communities. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 192a–192a. IEEE, 2005.

[199] Charlotte P Lee. Between chaos and routine: Boundary negotiating artifacts in collaboration. In *ECSCW 2005*, pages 387–406. Springer, 2005.

[200] Charlotte P Lee. Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing chaos in collaborative work. *Computer Supported Cooperative Work (CSCW)*, 16(3):307–339, 2007.

[201] Charlotte P Lee and Drew Paine. From the matrix to a model of coordinated action (MoCA): a conceptual framework of and for CSW. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 179–194. ACM, 2015.

[202] Minha Lee, Lily Frank, Femke Beute, Yvonne de Kort, and Wijnand Ijsselsteijn. Bots mind the social-technical gap. In *Proceedings of 15th European Conference on Computer-Supported Cooperative Work-Exploratory Papers*. European Society for Socially Embedded Technologies (EUSSET), 2017.

[203] Amanda Lenhart, Michele Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. Online harassment, digital abuse, and cyberstalking in america. Data & Society, January 2017.

[204] Lawrence Lessig. *Code version 2.0.* Basic Books, New York, [2nd ed.]. edition, 2006.

[205] Lawrence Lessig. *Code: And other laws of cyberspace*. ReadHowYouWant.com, 2009.

[206] Guo Li, Haiyi Zhu, Tun Lu, Xianghua Ding, and Ning Gu. Is it good to be like wikipedia?: Exploring the trade-offs of introducing collaborative editing model to Q&A sites. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1080–1091. ACM, 2015.

[207] Yue Lu, ChengXiang Zhai, and Neel Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th international Conference on World Wide Web*, pages 131–140. ACM, 2009.

[208] Kurt Luther, Nathan Hahn, Steven P Dow, and Aniket Kittur. Crowdlines: Supporting synthesis of diverse information sources through crowdsourced outlines. In *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.

[209] Wayne G Lutters and Mark S Ackerman. Achieving safety: a field study of boundary objects in aircraft technical support. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 266–275. ACM, 2002.

[210] Wayne G Lutters and Mark S Ackerman. Beyond boundary objects: collaborative reuse in aircraft technical support. *Computer Supported Cooperative Work (CSCW)*, 16(3):341–372, 2007.

[211] Gary S Lynn, Richard R Reilly, and Ali E Akgun. Knowledge management in new product teams: practices and outcomes. *IEEE transactions on Engineering Management*, 47(2):221–231, 2000.

[212] M Lynne Markus and Terry Connolly. Why cscw applications fail: Problems in the adoption of interdependent work tools. In *Proceedings of the 1990 ACM conference on Computer-supported cooperative work*, pages 371–380. ACM, 1990.

[213] Wendy E Mackay. More than just a communication system: diversity in the use of electronic mail. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, CSCW '98, pages 344–353, New York, NY, USA, 1998. ACM.

[214] Allan MacLean, Richard M Young, Victoria ME Bellotti, and Thomas P Moran. Questions, options, and criteria: Elements of design space analysis. *Human–computer interaction*, 6(3-4):201–250, 1991.

[215] Kaitlin Mahar, Amy X. Zhang, and David Karger. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 586:1–586:13, New York, NY, USA, 2018. ACM.

[216] Thomas W Malone, Robert Laubacher, and Chrysanthos Dellarocas. Harnessing crowds: Mapping the genome of collective intelligence. 2009.

[217] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. Design lessons from the fastest Q&A site in the west. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 2857–2866. ACM, 2011.

[218] Lena Mamykina, Drashko Nakikj, and Noemie Elhadad. Collective sensemaking in online health forums. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3217–3226, New York, NY, USA, 2015. ACM.

[219] Catherine C Marshall and Frank M Shipman. Spatial hypertext: designing for change. *Commun. ACM*, 38(8):88–97, 1995.

[220] Alice E Marwick and Robyn Caplan. Drinking male tears: language, the manosphere, and networked harassment. *Feminist Media Studies*, 18(4):543–559, 2018.

[221] Alice E Marwick et al. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133, 2011.

[222] Alice E Marwick and Ross W Miller. Online harassment, defamation, and hateful speech: A primer of the legal landscape. June 2014.

[223] Adrienne Massanari. # gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3):329–346, 2017.

[224] Sorin Adam Matei and Caius Dobrescu. Wikipedia's "neutral point of view": Settling conflict through ambiguity. *The Information Society*, 27(1):40–51, 2010.

[225] J Nathan Matias. Going dark: Social factors in collective action against platform operators in the reddit blackout. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 1138–1151. ACM, 2016.

[226] J. Nathan Matias. High impact questions and opportunities for online harassment research and action, August 2016.

[227] J Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. Reporting, reviewing, and responding to harassment on Twitter. May 2015.

[228] Uwe Matzat. Academic communication and internet discussion groups. *Social Networks*, 26(3):221–255, 2004.

[229] Willard McCarty. Humanist: Lessons from a global electronic seminar. *Computers and the Humanities*, 26(3):205–222, 1992.

[230] Amanda Menking and Ingrid Erickson. The heart work of Wikipedia: Gendered, emotional labor in the world's largest online encyclopedia. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 207–210. ACM, 2015.

[231] Microsoft. Microsoft teams. `https://teams.microsoft.com`.

[232] Matthew B Miles and A Michael Huberman. *Qualitative data analysis: An expanded sourcebook*. Sage, 1994.

[233] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. What do people ask their social networks, and why?: a survey study of status message Q&A behavior. In *CHI '10*, pages 1739–1748. ACM, 2010.

[234] Adam Mosseri. Bringing people closer together. `https://newsroom.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/`, Jan 2018.

[235] Xiangming Mu. Towards effective video annotation: An approach to automatically link notes with video content. *Computers & Education*, 55(4):1752–1763, 2010.

[236] Michael J. Muller and Daniel M. Gruen. Working together inside an emailbox. In *Proceedings ECSCW 2005*, ECSCW '05, pages 103–122. Springer, Dordrecht, 2005.

[237] Kevin Munger. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649, Sep 2017.

[238] Gabriel Murray and Giuseppe Carenini. Summarizing spoken and written conversations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 773–782. Association for Computational Linguistics, 2008.

[239] M. Naaman, J. Boase, and C. Lai. Is it really about me?: message content in social awareness streams. In *Proc. CSCW*, pages 189–192, 2010.

[240] Kevin K Nam and Mark S Ackerman. Arkose: reusing informal information from online discussions. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 137–146. ACM, 2007.

[241] Kanika Narang, Susan T. Dumais, Nick Craswell, Dan Liebling, and Qingyao Ai. Large-scale analysis of email search and organizational strategies. In *Proceedings of the 2017 Conference on Human Information Interaction&Retrieval*, CHIIR '17, pages 215–223, New York, NY, USA, 2017. ACM.

[242] Les Nelson, Rowan Nairn, Ed Chi, and Gregorio Convertino. Mail2tag: Augmenting email for sharing with implicit tag-based categorization. *Proc. CTS*, pages 23–30, 2011.

[243] Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. Springer, 2012.

[244] Ani Nenkova, Kathleen McKeown, et al. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233, 2011.

[245] Carman Neustaedter, A J. Bernheim Brush, Marc Smith, and Danyel Fisher. The social network and relationship finder: Social sorting for email triage. 01 2005.

[246] Beth Simone Noveck. *Wiki government: how technology can make government better, democracy stronger, and citizens more powerful.* Brookings Institution Press, 2009.

[247] Michael Nunez. Former facebook workers: We routinely suppressed conservative news. Gizmodo. Available: `https://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006`, May 2016.

[248] Jarkko Oikarinen and Darren Reed. Internet relay chat protocol. *IETF*, 1993. Available: `https://www.ietf.org/rfc/rfc1459.txt` [Last accessed: 2017-09-08], 1993.

[249] Gary M Olson and Judith S Olson. Distance matters. *Human–computer interaction*, 15(2-3):139–178, 2000.

[250] Judith S Olson, E Hofer, Nathan Bos, Ann Zimmerman, Gary M Olson, Daniel Cooney, and Ixchel Faniel. A theory of remote scientific collaboration. *Scientific collaboration on the internet*, pages 73–99, 2008.

[251] Wanda J Orlikowski. Sociomaterial practices: Exploring technology at work. *Organization studies*, 28(9):1435–1448, 2007.

[252] Elinor Ostrom. *Governing the commons.* Cambridge university press, 1990.

[253] Elinor Ostrom. Collective action and the evolution of social norms. *Journal of economic perspectives*, 14(3):137–158, 2000.

[254] Malcolm Otter and Hilary Johnson. Lost in hyperspace: metrics and mental models. *Interacting with Computers*, 13(1):1–40, 2000.

[255] Zizi Papacharissi. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society*, 6(2):259–283, 2004.

[256] Manoj Parameswaran and Andrew B Whinston. Research issues in social computing. *Journal of the Association for Information Systems*, 8(6):22, 2007.

[257] Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think.* Penguin, 2011.

[258] Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on Twitter. *CoRR*, abs/1706.01206, 2017.

[259] Namsu Park, Kerk Kee, and S Valenzuela. Being immersed in social networking environment. *CyberPsychology & Behavior*, 12(6):729–733, 2009.

[260] Soya Park, Amy X Zhang, Luke S Murray, and David R Karger. Opportunities for automating email processing: A need-finding study. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 374. ACM, 2019.

[261] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*, GROUP '16, pages 369–374, New York, NY, USA, 2016. ACM.

[262] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of LIWC2015. Technical report, 2015.

[263] Stephen T Peverly, Vivek Ramaswamy, Cindy Brown, James Sumowski, Moona Alidoost, and Joanna Garner. What predicts skill in lecture note taking? *Journal of Educational Psychology*, 99(1):167, 2007.

[264] Whitney Phillips. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Mit Press, 2015.

[265] Annie Piolat, Thierry Olive, and Ronald T Kellogg. Cognitive effort during note taking. *Applied Cognitive Psychology*, 19(3):291–312, 2005.

[266] Peter Pirolli and Stuart Card. Information foraging. *Psychological review*, 106(4):643, 1999.

[267] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, volume 5, pages 2–4, 2005.

[268] Jon Porter. Facebook to hire veteran journalists to curate upcoming news tab feature. The Verge. Available: `https://www.theverge.com/2019/8/20/20813833/facebook-hires-journalists-news-tab-publishers-fake-news-bias-top-stories`, Aug 2019.

[269] David G. Post. Anarchy, state, and the internet: An essay on law-making in cyberspace. *Journal of Online Law*, 1995:3–5, 1995.

[270] Jennifer Preece and Ben Shneiderman. The reader-to-leader framework: Motivating technology-mediated social participation. *AIS transactions on human-computer interaction*, 1(1):13–32, 2009.

[271] Jenny Preece. *Online communities: Designing usability and supporting sociability.* John Wiley & Sons, Inc., 2000.

[272] Reid Priedhorsky, Jilin Chen, Shyong Tony K Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 259–268. ACM, 2007.

[273] Robert D Putnam, Robert Leonardi, and Raffaella Y Nanetti. *Making democracy work: Civic traditions in modern Italy.* Princeton university press, 1994.

[274] Anabel Quan-Haase, Joseph Cothrel, and Barry Wellman. Instant messaging for collaboration: A case study of a high-tech firm. *Journal of Computer-Mediated Communication*, 10(4):00–00, 2005.

[275] Quora. `https://quora.com`.

[276] Owen Rambow, Lokesh Shrestha, John Chen, and Chirsty Lauridsen. Summarizing email threads. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 105–108. Association for Computational Linguistics, 2004.

[277] Reddit. `https://reddit.com`.

[278] Elizabeth Reid. Electropolis: Communication and community on internet relay chat, 1991.

[279] Yuqing Ren, Robert Kraut, and Sara Kiesler. Applying common identity and bond theory to design of online communities. *Organization studies*, 28(3):377–408, 2007.

[280] Evan F Risko, Tom Foulsham, Shane Dawson, and Alan Kingstone. The collaborative lecture annotation system (clas): A new tool for distributed learning. *IEEE Transactions on Learning Technologies*, 6(1):4–13, 2013.

[281] Sarah T Roberts. *Behind the Screen: Content Moderation in the Shadows of Social Media.* Yale University Press, 2019.

[282] Alejandra Rojo and Ronald Ragsdale. A process perspective on participation in scholarly electronic forums. *Science communication*, 18(4):320–341, 1997.

[283] Adam Rule, Ian Drosos, Aurélien Tabard, and James D Hollan. Aiding collaborative reuse of computational notebooks with annotated cell folding. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):150, 2018.

[284] Adam Rule, Aurélien Tabard, and James D Hollan. Exploration and explanation in computational notebooks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 32. ACM, 2018.

[285] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[286] Maya Sappelli, Gabriella Pasi, Suzan Verberne, Maaike de Boer, and Wessel Kraaij. Assessing e-mail intent and tasks in e-mail messages. *Information Sciences*, 358:1–17, 2016.

[287] Kjeld Schmidt and Carla Simonee. Coordination mechanisms: Towards a conceptual foundation of cscw systems design. *Computer Supported Cooperative Work (CSCW)*, 5(2-3):155–200, 1996.

[288] Jodi Schneider, John G Breslin, and Alexandre Passant. A content analysis: How wikipedia talk pages are used. WebSci, 2010.

[289] Jodi Schneider, Alexandre Passant, and John G Breslin. Understanding and improving wikipedia article discussion spaces. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 808–813. ACM, 2011.

[290] Douglas Schuler and Aki Namioka. *Participatory design: Principles and practices.* CRC Press, 1993.

[291] John R Searle. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press, 1969.

[292] Max Seddon. Documents show how Russia's troll army hit america. Buzzfeed. Available: `https://www.buzzfeednews.com/article/maxseddon/documents-show-how-russias-troll-army-hit-america`, June 2014.

[293] Joseph Seering, Robert Kraut, and Laura Dabbish. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 111–125, New York, NY, USA, 2017. ACM.

[294] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, page 1461444818821316, 2019.

[295] Amy Shapiro and Dale Niederhauser. Learning from hypertext: Research issues and findings. *Handbook of Research on Educational Communications and Technology*, 2:605–620, 2004.

[296] Bayan Abu Shawar and Eric Atwell. Chatbots: are they really useful? In *LDV Forum*, volume 22, pages 29–49, 2007.

[297] Bayan Abu Shawar and Eric Steven Atwell. Using corpora in machine-learning chatbot systems. *International journal of corpus linguistics*, 10(4):489–516, 2005.

[298] Tamara Shepherd, Alison Harvey, Tim Jordan, Sam Srauy, and Kate Miltner. Histories of hating. *Social Media + Society*, 1(2), 2015.

[299] Emad Shihab, Zhen Ming Jiang, and Ahmed E Hassan. On the use of internet relay chat (irc) meetings by developers of the gnome gtk+ project. In *Mining Software Repositories, 2009. MSR'09. 6th IEEE International Working Conference on*, pages 107–110. IEEE, 2009.

[300] Frank M Shipman and Catherine C Marshall. Formality considered harmful: Experiences, emerging themes, and directions on the use of formal representations in interactive systems. *Computer Supported Cooperative Work (CSCW)*, 8(4):333–352, 1999.

[301] Frank M Shipman III and Raymond McCall. Supporting knowledge-base evolution with incremental formalization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 285–291. Citeseer, 1994.

[302] Clay Shirky. *Here comes everybody: The power of organizing without organizations*. Penguin, 2008.

[303] Yih-Chearng Shiue, Chao-Min Chiu, and Chen-Chi Chang. Exploring and mitigating social loafing in online communities. *Computers in Human Behavior*, 26(4):768–777, 2010.

[304] Herbert A Simon. Designing organizations for an information-rich world. *International Library of Critical Writings in Economics*, 70:187–202, 1996.

[305] Nikash Singh, Martin Tomitsch, and Mary Lou Maher. Understanding the management and need for awareness of temporal information in email. In *Proceedings of the 2006 international workshop on Mining software repositories*, MSR '06, pages 137–143, New York, NY, USA, 2006. ACM.

[306] Slack. https://slack.com.

[307] Slack. Focus on the important things with highlights in slack. *Slack*, 14 June 2017. Available: https://slackhq.com/focus-on-the-important-things-with-highlights-in-slack-5e30024502cd [Last accessed: 2017-09-15].

[308] Manya Sleeper, Rebecca Balebako, Sauvik Das, Amber McConahy, Jason Wiese, and Lorrie Cranor. The post that wasn't. *Proc. CSCW*, pages 793–802, 2013.

[309] Aaron Smith and Maeve Duggan. Crossing the line: What counts as online harassment? The Pew Research Center, January 2018.

[310] Marc Smith, Jonathan J Cadiz, and Byron Burkhalter. Conversation trees and threaded chats. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 97–105. ACM, 2000.

[311] Andrew Smock, Nicole Ellison, Cliff Lampe, and Donghee Wohn. Facebook as a toolkit. *Computers in Human Behavior*, 27(6):2322–2329, 2011.

[312] Tiziano Squartini, Francesco Picciolo, Franco Ruzzenenti, and Diego Garlaschelli. Reciprocity of weighted networks. *Scientific reports*, 3:2729, 2013.

[313] StackOverflow. `http://stackoverflow.com`.

[314] Susan Leigh Star and James R Griesemer. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in berkeley's museum of vertebrate zoology, 1907-39. *Social studies of science*, 19(3):387–420, 1989.

[315] Robert A Stebbins. Serious leisure: A conceptual statement. *Pacific sociological review*, 25(2):251–272, 1982.

[316] Hans Stiegler and Menno DT de Jong. Facilitating personal deliberation online: Immediate effects of two considerit variations. *Computers in human behavior*, 51:461–469, 2015.

[317] Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

[318] Anselm Strauss. The articulation of project work: An organizational process. *Sociological Quarterly*, 29(2):163–178, 1988.

[319] Besiki Stvilia, Michael B Twidale, Linda C Smith, and Les Gasser. Information quality work organization in wikipedia. *Journal of the Association for Information Science and Technology*, 59(6):983–1001, 2008.

[320] Lucy Suchman. Supporting articulation work. *Computerization and controversy: Value conflicts and social choices*, 2:407–423, 1996.

[321] Lucy Suchman. Making work visible. In *The New Production of Users*, pages 143–153. Routledge, 2016.

[322] Róbert Sumi, Taha Yasseri, et al. Edit wars in wikipedia. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 724–727. IEEE, 2011.

[323] Olof Sundin. Janitors of knowledge: constructing knowledge in the everyday life of wikipedia editors. *Journal of documentation*, 67(5):840–862, 2011.

[324] James Surowiecki. Wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations, 2004.

[325] Saiganesh Swaminathan, Raymond Fok, Fanglin Chen, Ting-Hao Kenneth Huang, Irene Lin, Rohan Jadvani, Walter S. Lasecki, and Jeffrey P. Bigham.

Wearmail: On-the-go access to information in your email with a privacy-preserving human computation workflow. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, UIST '17, pages 807–815, New York, NY, USA, 2017. ACM.

[326] Agnieszka Matysiak Szóstek. 'dealing with my emails': Latent user needs in email management. *Computers in Human Behavior*, 27(2):723–729, 2011.

[327] Henri Tajfel. Social identity and intergroup behaviour. *Information (International Social Science Council)*, 13(2):65–93, 1974.

[328] Sanna Talja, Reijo Savolainen, and Hanni Maula. Field differences in the use and perceived usefulness of scholarly mailing lists. *Information research*, 10(1), 2004.

[329] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624. International World Wide Web Conferences Steering Committee, 2016.

[330] Dario Taraborelli and Giovanni Luca Ciampaglia. Beyond notability. collective deliberation on content inclusion in wikipedia. In *Self-Adaptive and Self-Organizing Systems Workshop (SASOW), 2010 Fourth IEEE International Conference on*, pages 122–125. IEEE, 2010.

[331] Jaime Teevan. The future of microwork. *XRDS: Crossroads*, 23(2):26–29, 2016.

[332] Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 415–422, New York, NY, USA, 2004. ACM.

[333] Jaime Teevan, Shamsi T. Iqbal, and Curtis von Veh. Supporting collaborative writing with microtasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2657–2668, New York, NY, USA, 2016. ACM.

[334] Jaime Teevan, Daniel J. Liebling, and Walter S. Lasecki. Selfsourcing personal tasks. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, pages 2527–2532. ACM, 2014.

[335] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. Understanding chatbot-mediated task management. ACM, 2018.

[336] Susanne Trauzettel-Klosinski and Klaus Dietz. Standardized assessment of reading performance: The new international reading speed texts ireststandardized

assessment of reading performance. *Investigative Ophthalmology & Visual Science*, 53(9):5452–5461, 2012.

[337] Zeynep Tufekci. *Twitter and tear gas: The power and fragility of networked protest.* Yale University Press, 2017.

[338] Edward R Tufte. *Beautiful evidence*, volume 1. Graphics Press Cheshire, CT, 2006.

[339] Max G Van Kleek, Michael Bernstein, Katrina Panovich, Gregory G Vargas, David R Karger, and MC Schraefel. Note to self: examining personal information keeping in a lightweight note-taking tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1477–1480. ACM, 2009.

[340] Bogdan Vasilescu, Alexander Serebrenik, Prem Devanbu, and Vladimir Filkov. How social q&a sites are changing knowledge sharing in open source software communities. In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing*, pages 342–354. ACM, 2014.

[341] Gina Danielle Venolia and Carman Neustaedter. Understanding sequence and reply relationships within email conversations: a mixed-model visualization. CSCW '03, pages 361–368. ACM, 2003.

[342] Vasilis Verroios and Michael S Bernstein. Context trees: Crowdsourcing global understanding from local views. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[343] Fernanda B Viégas, Scott Golder, and Judith Donath. Visualizing email content: portraying relationships from conversational histories. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 979–988. ACM, 2006.

[344] Fernanda B Viegas, Martin Wattenberg, Jesse Kriss, and Frank Van Ham. Talk before you type: Coordination in Wikipedia. In *System sciences, 2007. HICSS 2007. 40th annual Hawaii international conference on*, pages 78–78. IEEE, 2007.

[345] Jessica Vitak. The impact of context collapse and privacy on social network site disclosures. *Journal of Broadcasting & Electronic Media*, 56(4):451–470, 2012.

[346] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1231–1245, New York, NY, USA, 2017. ACM.

[347] Luis Von Ahn. Human computation. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 1–2. IEEE Computer Society, 2008.

[348] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledge base. 2014.

[349] W3C. Zakim. 7 September 2015. Available: `https://www.w3.org/2001/12/zakim-irc-bot.html` [Last accessed: 2018-04-18].

[350] Jessica Wang. A system for bridging the ideological divide by establishing a moral framework for news consumption, 2017.

[351] Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policy making. Coucil of Europe Report DGI(2017)09, Council of Europe, October 2017.

[352] Charlie Warzel. "A honeypot for assholes": Inside Twitter's 10-year failure to stop harassment. Buzzfeed, `https://www.buzzfeed.com/charliewarzel/a-honeypot-for-assholes-inside-twitters-10-year-failure-to-s`, Last access: September 8, 2017, August 2016.

[353] Charlie Warzel. Twitter is still dismissing harassment reports and frustrating victims. Buzzfeed, Available: `https://www.buzzfeed.com/charliewarzel/twitter-is-still-dismissing-harassment-reports-and` [Last accessed: September 8, 2017], July 2017.

[354] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. pages 78–84, 2017.

[355] Joseph Weizenbaum. Eliza - a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[356] Steve Whittaker, Victoria Bellotti, and Jacek Gwizdka. Email in personal information management. *Communications of the ACM*, 49(1):68–73, 2006.

[357] Steve Whittaker and Candace Sidner. Email overload: exploring personal information management of email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 793–802, 1996.

[358] WHOA. WHOA (working to halt online abuse) comparison statistics 2000-2013. Available: `http://www.haltabuse.org/resources/stats/Cumulative2000-2013.pdf`, 2013.

[359] Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497. Springer, 2005.

[360] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.

[361] Wikipedia. `https://en.wikipedia.org`.

[362] Wikipedia. Wikipedia:flow. `https://en.wikipedia.org/wiki/Wikipedia:Flow`.

[363] Raelene Wilding. "Virtual" intimacies? families communicating across transnational contexts. *Global networks*, 6(2):125–142, 2006.

[364] Janis Wolak, Kimberly J Mitchell, and David Finkelhor. Does online harassment constitute bullying? an exploration of online harassment by known peers and online-only contacts. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine*, 41(6):S51–S58, 2007.

[365] Gadi Wolfsfeld, Elad Segev, and Tamir Sheafer. Social media and the arab spring: Politics comes first. *The International Journal of Press/Politics*, 18(2):115–137, 2013.

[366] Joon Ian Wong, Dave Gershgorn, and Mike Murphy. Facebook is trying to get rid of bias in trending news by getting rid of humans, Aug 2016.

[367] David R Woolley. Plato: The emergence of online community, 1994.

[368] Lori Wright. Expand your collaboration with guest access in microsoft teams. *Microsoft Office Blog*, 11 September 2017. Available: `https://blogs.office.com/en-us/2017/09/11/expand-your-collaboration-with-guest-access-in-microsoft-teams/` [Last accessed: 2017-09-14].

[369] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1391–1399, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[370] Wensi Xi, Jesper Lind, and Eric Brill. Learning effective ranking functions for newsgroup search. In *SIGIR '04*, pages 394–401. ACM, 2004.

[371] Liu Yang, Susan T Dumais, Paul N Benne, and Ahmed Hassan Awadallah. Characterizing and predicting enterprise email reply behavior. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2017, Tokyo, Japan*, 2017.

[372] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. Dynamics of conflicts in wikipedia. *PLoS ONE*, 7(6):e38869, 2012.

[373] Alexander Seeshing Yeung, Putai Jin, and John Sweller. Cognitive load and learner expertise: Split-attention and redundancy effects in reading with explanatory notes. *Contemporary educational psychology*, 23(1):1–21, 1998.

[374] Shinjae Yoo, Yiming Yang, and Jaime Carbonell. Modeling personalized email prioritization: classification-based and regression-based approaches. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 729–738, New York, NY, USA, 2011. ACM.

[375] Kathryn Sue Young, Julia T Wood, Gerald M Phillips, and Douglas J Pedersen. *Group discussion: A practical guide to participation and leadership*. Waveland Press, 2006.

[376] David M Zajic, Bonnie J Dorr, and Jimmy Lin. Single-document and multi-document summarization techniques for email threads using sentence compression. *Information Processing & Management*, 44(4):1600–1610, 2008.

[377] Amy X. Zhang, Mark S. Ackerman, and David R. Karger. Mailing lists: Why are they still here, what's wrong with them, and how can we fix them? In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 4009–4018, New York, NY, USA, 2015. ACM.

[378] Amy X. Zhang, Joshua Blum, and David R. Karger. Opportunities and challenges around a tool for social and public web activity tracking. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 913–925, New York, NY, USA, 2016. ACM.

[379] Amy X. Zhang and Scott Counts. Modeling ideology and predicting policy change with social media: Case of same-sex marriage. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2603–2612, New York, NY, USA, 2015. ACM.

[380] Amy X. Zhang and Scott Counts. Gender and ideology in the spread of anti-abortion policy. In Proceedings of CHI 2016*: ACM Conference on Human Factors in Computing Systems*, San Jose, CA, 2016.

[381] Amy X. Zhang and Justin Cranshaw. Making sense of group chat through collaborative tagging and summarization. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*. ACM, 2018.

[382] Amy X. Zhang, Bryan Culbertson, and Praveen Paritosh. Characterizing online discussion using coarse discourse sequences. In *Proceedings of the 11th International AAAI Conference on Weblogs and Social Media*, ICWSM '17, 2017.

[383] Amy X Zhang, Michele Igo, Marc Facciotti, and David Karger. Using student annotated hashtags and emojis to collect nuanced affective states. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 319–322. ACM, 2017.

[384] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer

Lee, Martin Robbins, et al. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, pages 603–612. International World Wide Web Conferences Steering Committee, 2018.

[385] Amy X Zhang, Lea Verou, and David R Karger. Wikum: Bridging discussion forums and wikis using recursive summarization. In *CSCW*, pages 2082–2096, 2017.

[386] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. Value-sensitive algorithm design: method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):194, 2018.

[387] Mark Zuckerberg. A blueprint for content governance and enforcement. `https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/`, Nov 2018.

[388] Sacha Zyto, David Karger, Mark Ackerman, and Sanjoy Mahajan. Successful classroom deployment of a social document annotation system. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 1883–1892. ACM, 2012.