Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation

Kaitlin Mahar
MIT CSAIL
Cambridge, MA, USA
kmahar@mit.edu

Amy X. Zhang
MIT CSAIL
Cambridge, MA, USA
axz@mit.edu

David Karger
MIT CSAIL
Cambridge, MA, USA
karger@mit.edu

ABSTRACT

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.

Author Keywords

online harassment; email; moderation; private messages; friendsourcing; crowdsourcing; social media

ACM Classification Keywords

H.5.3. Group and Organization Interfaces: Asynchronous interaction; Web-based interaction

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 [23] and the Pew Research Center [8], 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 [16, 35]. Recently, researchers have built machine learning models to detect harassment [6, 20, 39], but caution that such models should be used in tandem with human moderators [1], due to the possibility of deception [18] and presence of bias in training data [3]. Indeed, paid human moderators already make up many of the reporting pipelines for platforms [26], but they still often fail to understand the nuances of people's experiences [4] and make opaque or inconsistent decisions [29, 36].

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

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

© 2018 ACM. ISBN 978-1-4503-5620-6/18/04. . . $15.00

DOI: https://doi.org/10.1145/3173574.3174160

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.

RELATED WORK

Online Harassment Research

There has been a great deal of work characterizing online harassment as a significant problem affecting many internet users [8, 23], with certain groups such as young adults [33, 38], women [11, 31, 33, 34], and those who identify as LGBTQ [23] 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 [23]. As a result of harassment, many recipients simply withdraw from public online spaces [11, 34] or self-censor their content online [23]. Researchers and internet activists have studied or called for better processes to deal with harassment on various platforms [16, 26, 29]. Other researchers examine government policy on online harassment, finding it ineffective [24]. Researchers have also suggested design interventions for platforms to undertake, resulting from content analysis [30], interviews and surveys [34], and design sessions [2] with harassment recipients.

Technical Solutions for Combating Harassment

Researchers and platforms have built technical solutions to combat unwanted messages, beginning with address blocklists and text-based email filters in the early days of the internet. Most social media platforms have also incorporated these tools. In more recent years, some researchers have built classifiers to detect harassing, trolling, or otherwise toxic content, using hand-labeled data [12, 28, 39] or content from existing communities [6]. Researchers have also worked to release data [14] and to better define subtasks within the overall space [20, 37]. However, researchers have also qualified this work, warning that such models have documented errors and should not be used without human oversight [1]. Studying existing models, researchers found they could be easily deceived into misclassifying abusive messages [18]. Others found significant differences in data labeling performed by women and men [3], suggesting automated systems can inherit the biases of their data. Additionally, researchers suggest that wide differences in norms between communities may make labeled data from one community untransferable to another [3]. Given the criticisms, purely automated approaches to combat harassment are not a complete solution in the near-term.

Community-Based Systems for Combating Harassment

By building on prior research methods and findings [10, 25], 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 [13]. Some of these tools include BlockTogether [17] and Good Game Auto Blocker [15], where users collaborate on shared blocklists of harassing Twitter accounts. Other community-based efforts include projects such as Hollaback! that elevate victims' stories [7], and systems such as HeartMob that provide a network of volunteers to support, provide validation for, and take action on behalf of harassment recipients [4]. 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 [2] by learning from harassment recipients' existing strategies to then design a tool to augment those strategies.

Collaborative Systems for Message Management

Finally, we draw from research on systems for collaborative management and moderation of messages delivered to and from an individual. Our group explored email usage in mailing lists, finding use cases for friendsourced moderation of one's outgoing email to overcome anxieties about posting to a public list [40]. Other researchers have studied the use of crowdsourced workers to provide personal email management services. Kokkalis et al. use remote microtask workers to extract tasks and manage email overload [21, 22], finding that over time users became more comfortable with strangers seeing their emails. We build on this work by examining friend moderators, who have many advantages over strangers—they are personally motivated to help and have a deeper understanding of context. Privacy considerations for friends are also significantly different than those for strangers.

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, text message	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 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, Text message	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]	Text message	SMS spoofing - both received messages, and messages sent pretending to be this person. Unclear who is the harasser.	1+	(No response)
Public Figure [Pub1]	Twitter, Email	Large volume of continual harassment, including greater waves due to public appearances. Personal attacks and death threats.	50+	(No response)

Table 1. 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).

EXPERIENCES, PREFERENCES, AND STRATEGIES

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 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 [32]. 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 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], highlighting the oftentimes persistent nature of harassers.

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 [23]. 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 [24].

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 [23]. 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.

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.*" 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

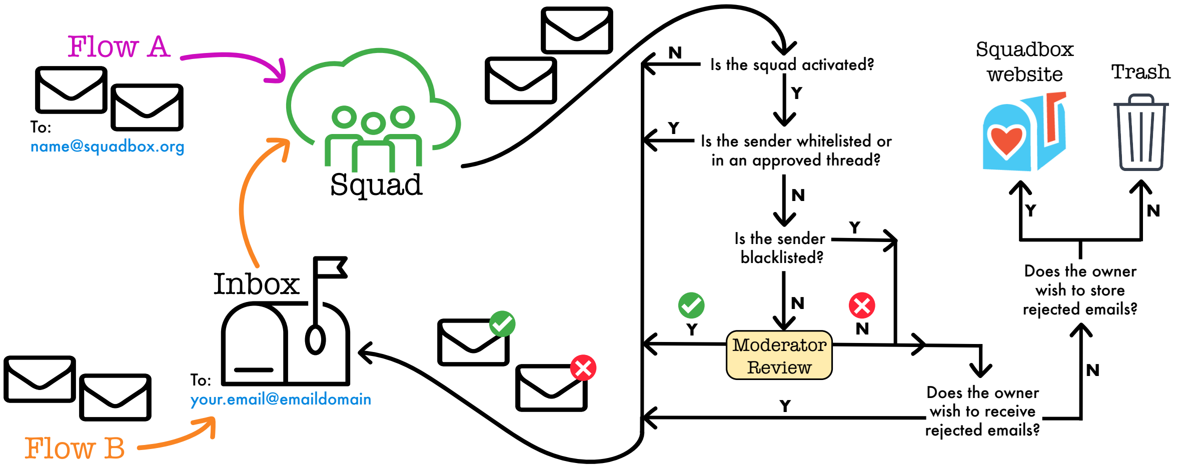


Figure 1. Diagram of the flow of emails through Squadbox, including Flow A, which allows users to have a public moderated account, and Flow B, which allows users to get their current account moderated. From there, various settings define whether emails get moderated and where they go.

harassment could have on moderators, there is still risk of secondary trauma, as content moderators for platforms have described [5]. 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], echoing work on the effectiveness of certain identities in bystander intervention [27]. However, Pub1 felt that certain insults targeted at an identity might not be recognized by people outside of that identity unless they were trained.

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**).

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 1, followed by features and implementation of the system. From here onward, we use the term "owner" to refer to the person who is having their emails moderated.

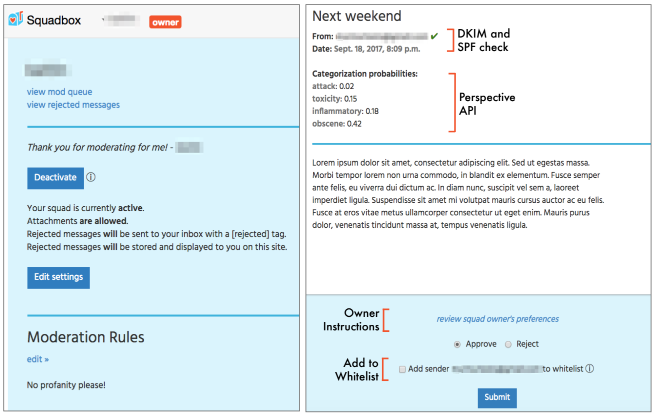
User Scenarios

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 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).

[1] Squadbox: http://squadbox.org



[redacted]

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.

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.

● Approve ○ Reject

□ Add sender [redacted] to whitelist ⓘ

Figure 2. On the left, an owner's view of the information page for their squad. On the right, a moderation page for the moderator.

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 [19] 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 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.

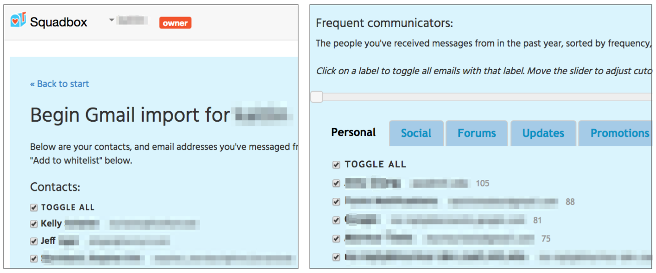
Customized Instructions: As people have different ideas about what is harassment [32] or have different actions they want moderators to take, we allow owners to give instructions to their moderators via a freeform text box (G1).

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 [19]. 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).


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

Personal | Social | Forums | Updates | Promotions

☑ TOGGLE ALL
☑ [redacted] 105
☑ [redacted] 88
☑ [redacted] 81
☑ [redacted] 75
☑ [redacted]

Figure 3. Squadbox generates whitelist suggestions from owner's Gmail contacts.

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 inserted at the top of the email if the owner has chosen to have rejected messages delivered.

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 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 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.

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.

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,

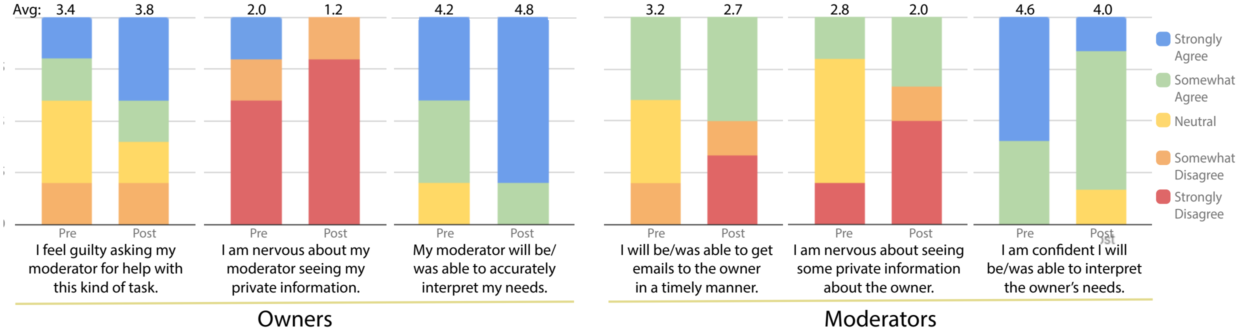


Figure 4. Comparison of agreement (where 1=Strongly Disagree, 5=Strongly Agree) with statements before and after the field study.

Squad	WL Size	% Accept	% Reject	Total Volume
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
Average	326.8	47.4	52.6	60

Table 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.

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 identity obfuscation, for fear that moderators may try to retaliate against harassers.

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.

Field Study Results

The whitelist/blacklist feature was an effective way to separate out potentially unwanted messages. As shown in Table 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 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 4. In addition, owners felt more guilty over the study.

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.

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 [22] 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 [4]. However, we would need to set checks to protect against harassers seeking to infiltrate the system.

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.

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 [9]. 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.

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.

REFERENCES

1. CJ Adams and Lucas Dixon. 2017. Better discussions with imperfect models. (11 September 2017). Retrieved January 3, 2018 from https://medium.com/the-false-positive/better-discussions-with-imperfect-models-91558235d442

2. Zahra Ashktorab and Jessica Vitak. 2016. 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)*. ACM, New York, NY, USA, 3895–3905. DOI: http://dx.doi.org/10.1145/2858036.2858548

3. Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. 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*. Springer, Springer International Publishing, 405–415. DOI: http://dx.doi.org/10.1007/978-3-319-67256-4_32

4. Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 24 (Dec. 2017), 19 pages. DOI: http://dx.doi.org/10.1145/3134659

5. Michael L. Bourke and Sarah W. Craun. 2014. Secondary Traumatic Stress Among Internet Crimes Against Children Task Force Personnel: Impact, Risk Factors, and Coping Strategies. *Sexual Abuse* 26, 6 (2014), 586–609. DOI:http://dx.doi.org/10.1177/1079063213509411

6. Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. 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)*. ACM, New York, NY, USA, 3175–3187. DOI: http://dx.doi.org/10.1145/3025453.3026018

7. Jill P. Dimond, Michaelanne Dye, Daphne Larose, and Amy S. Bruckman. 2013. Hollaback!: The Role of Storytelling Online in a Social Movement Organization. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 477–490. DOI: http://dx.doi.org/10.1145/2441776.2441831

8. Maeve Duggan. 2017. Online Harassment 2017. The Pew Research Center. (11 July 2017). Retrieved September 8, 2017 from http://www.pewinternet.org/2017/07/11/online-harassment-2017/

9. Steven Englehardt, Jeffrey Han, and Arvind Narayanan. 2018. I never signed up for this! Privacy implications of email tracking. *Proceedings on Privacy Enhancing Technologies* 1 (2018), 109–126. DOI: http://dx.doi.org/10.1515/popets-2018-0006

10. Robert Faris, Amar Ashar, Urs Gasser, and Daisy Joo. 2016. Understanding Harmful Speech Online. Berkman Klein Center Research Publication 2016-21, (8 December 2016). DOI:http://dx.doi.org/10.2139/ssrn.2882824

11. Jesse Fox and Wai Yen Tang. 2017. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media & Society* 19, 8 (2017), 1290–1307. DOI: http://dx.doi.org/10.1177/1461444816635778

12. Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-Speech. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, 85–90. http://aclweb.org/anthology/W17-3013

13. R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (21 March 2016), 787–803. https://ssrn.com/abstract=2761503

14. 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. 2017. A Large Labeled Corpus for Online Harassment Research. In *Proceedings of the 2017 ACM on Web Science Conference (WebSci '17)*. ACM, New York, NY, USA, 229–233. DOI: http://dx.doi.org/10.1145/3091478.3091509

15. Randi Lee Harper. 2014. Good Game Auto Blocker. (2014). Retrieved September 8, 2017 from https://github.com/freebsdgirl/ggautoblocker

16. Randi Lee Harper. 2016. Putting out the Twitter trashfire. Art + Marketing. (13 February 2016). Retrieved September 8, 2017 from https://artplusmarketing.com/putting-out-the-twitter-trashfire-3ac6cb1af3e

17. Jacob Hoffman-Andrews. 2017. BlockTogether. (2017). Retrieved September 8, 2017 from https://blocktogether.org/

18. Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *CoRR* abs/1702.08138 (2017). http://arxiv.org/abs/1702.08138

19. Google Jigsaw. 2017. Perspective API. (2017). Retrieved September 8, 2017 from https://www.perspectiveapi.com/

20. George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. 2017. Technology Solutions to Combat Online Harassment. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, 73–77. http://aclweb.org/anthology/W17-3013

21. Nicolas Kokkalis, Chengdiao Fan, Johannes Roith, Michael S. Bernstein, and Scott Klemmer. 2017. MyriadHub: Efficiently Scaling Personalized Email Conversations with Valet Crowdsourcing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 73–84. DOI: http://dx.doi.org/10.1145/3025453.3025954

22. Nicolas Kokkalis, Thomas Köhn, Carl Pfeiffer, Dima Chornyi, Michael S. Bernstein, and Scott R. Klemmer. 2013. EmailValet: Managing Email Overload Through Private, Accountable Crowdsourcing. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1291–1300. DOI: http://dx.doi.org/10.1145/2441776.2441922

23. Amanda Lenhart, Michele Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. 2017. Online Harassment, Digital Abuse, and Cyberstalking in America. Data & Society. (18 January 2017). Retrieved September 8, 2017 from https://datasociety.net/blog/2017/01/18/online-harassment-digital-abuse

24. Alice E Marwick and Ross W Miller. 2014. Online harassment, defamation, and hateful speech: A primer of the legal landscape. Fordham Center on Law and Information Policy Report, (10 June 2014). http://ssrn.com/abstract=2447904

25. J. Nathan Matias. 2016. High Impact Questions and Opportunities for Online Harassment Research and Action. (August 2016). Retrieved September 8, 2017 from https://civic.mit.edu/sites/civic.mit.edu/files/OnlineHarassmentWorkshopReport-08.2016.pdf

26. J Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. 2015. Reporting, reviewing, and responding to harassment on Twitter. Women, Action, and the Media, (13 May 2015). http://womenactionmedia.org/twitter-report/

27. Kevin Munger. 2017. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior* 39, 3 (01 Sep 2017), 629–649. DOI: http://dx.doi.org/10.1007/s11109-016-9373-5

28. Ji Ho Park and Pascale Fung. 2017. One-step and Two-step Classification for Abusive Language Detection on Twitter. *CoRR* abs/1706.01206 (2017). http://arxiv.org/abs/1706.01206

29. Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 19th International Conference on Supporting Group Work (GROUP '16)*. ACM, New York, NY, USA, 369–374. DOI: http://dx.doi.org/10.1145/2957276.2957297

30. Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. 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)*. ACM, New York, NY, USA, 111–125. DOI: http://dx.doi.org/10.1145/2998181.2998277

31. Tamara Shepherd, Alison Harvey, Tim Jordan, Sam Srauy, and Kate Miltner. 2015. Histories of Hating. *Social Media + Society* 1, 2 (2015). DOI: http://dx.doi.org/10.1177/2056305115603997

32. Aaron Smith and Maeve Duggan. 2018. Crossing the Line: What Counts as Online Harassment? The Pew Research Center. (4 January 2018). Retrieved January 8, 2018 from http://www.pewinternet.org/2018/01/04/crossing-the-line-what-counts-as-online-harassment/

33. Working to Halt Online Abuse. 2013. WHOA Comparison Statistics 2000-2013. (2013). Retrieved September 8, 2017 from http://www.haltabuse.org/resources/stats/Cumulative2000-2013.pdf

34. Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. 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)*. ACM, New York, NY, USA, 1231–1245. DOI: http://dx.doi.org/10.1145/2998181.2998337

35. Charlie Warzel. 2016. "A Honeypot For Assholes": Inside Twitter's 10-Year Failure To Stop Harassment. (11 August 2016). Retrieved September 8, 2017 from https://www.buzzfeed.com/charliewarzel/a-honeypot-for-assholes-inside-twitters-10-year-failure-to-s

36. Charlie Warzel. 2017. Twitter Is Still Dismissing Harassment Reports And Frustrating Victims. Buzzfeed. (17 July 2017). Retrieved September 8, 2017 from https://www.buzzfeed.com/charliewarzel/twitter-is-still-dismissing-harassment-reports-and

37. Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. (2017), 78–84. http://aclweb.org/anthology/W17-3012

38. Janis Wolak, Kimberly J Mitchell, and David Finkelhor. 2007. 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 (2007), S51–S58. DOI: http://dx.doi.org/10.1016/j.jadohealth.2007.08.019

39. Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1391–1399. DOI: http://dx.doi.org/10.1145/3038912.3052591

40. Amy X. Zhang, Mark S. Ackerman, and David R. Karger. 2015. 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)*. ACM, New York, NY, USA, 4009–4018. DOI: http://dx.doi.org/10.1145/2702123.2702194