# Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation

by

Kaitlin E. Mahar

S.B., Massachusetts Institute of Technology (2016)

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

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2017



Author .....
Department of Electrical Engineering and Computer Science
September 8, 2017

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

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

# Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation

by

Kaitlin E. Mahar

Submitted to the Department of Electrical Engineering and Computer Science
on September 8, 2017, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

## Abstract

Online harassment is both a social and technical problem, and has yet to be successfully solved. In order to better understand the issue and the challenges surrounding it, we conducted interviews with eighteen recipients of online harassment to understand their current strategies for coping, finding that they often resorted to asking friends for help. Inspired by these findings, we built Squadbox.

Squadbox is a tool to help people experiencing email harassment by allowing them organize a "squad" of friend moderators to protect them from harassing messages. The moderators intercept emails 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. This thesis details the interview findings and their design implications, the implementation of the Squadbox prototype, and the findings from the field study.

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

# Acknowledgments

I am extremely grateful that I've had the opportunity to work with my advisor, David Karger, and his Ph. D. student Amy Zhang over the last three semesters. In writing this thesis, I've realized just how much they've taught me.

David, you have provided excellent guidance throughout the course of this project. You helped me develop a very useful mental framework for thinking about our work and its future potential, and I've really enjoyed having long brainstorming sessions with you and Amy. It's easy for me to get bogged down in the details when I'm focused on coding, and I appreciate your focus on the bigger picture.

Amy, you are a wonderful mentor and I cannot thank you enough. You've helped me chase down the strangest of encoding bugs, sent me countless related papers, and given me great advice. Your enthusiasm for your work is contagious, and helped keep me motivated throughout the year. Thank you for teaching me so much about how to do research.

To my family, friends, and boyfriend, thank you so much for your love and support. I could not have made it through this year without having such a strong support system outside of school.

I would also like to thank our UROP Oliver Dunkley, who made major contributions to building the Squadbox prototype, and the rest of the Haystack Group for their ideas and encouragement.

And lastly, this research would not have been possible without the interview and study subjects who volunteered their time and ideas to help with the project. Thank you for your willingness to speak with us about such deeply personal experiences.

# Contents

**1 Introduction** **15**

**2 Related Work** **19**

2.1 Online Harassment Research . . . . . 19

2.2 Technical Solutions for Combating Harassment . . . . . 20

2.3 Community-Based Systems for Combating Harassment . . . . . 20

2.4 Collaborative Systems for Message Management . . . . . 21

**3 User Interviews** **23**

3.1 Methodology . . . . . 23

3.2 Interview Subjects . . . . . 24

3.3 Findings . . . . . 24

3.3.1 How Subjects Defined Harassment . . . . . 26

3.3.2 Consequences of Harassment . . . . . 27

3.3.3 Actions Taken in Response to Harassment . . . . . 31

3.3.4 Desired Platform Changes and Features . . . . . 36

3.3.5 Desired Assistance from Other People . . . . . 40

3.3.6 Friendsourced Moderation Feedback . . . . . 41

**4 System Design and Implementation** **63**

4.1 User Needs and Design Goals . . . . . 63

4.2 User Scenarios . . . . . 65

4.2.1 Flow A: Squadbox as a public contact address . . . . . 65

4.2.2 Flow B: Squadbox with an existing email account . . . . . 66
4.3 Squadbox Features . . . . . 66
4.3.1 Features for Reducing Moderator Load and Increasing Privacy 66
4.3.2 Features for Reducing Secondary Trauma to Moderators . . . 68
4.3.3 Features for Giving Moderators Context and Information . . . 69
4.3.4 Features for Giving Owners Customization Capabilities . . . . 71
4.3.5 Moderator Tags . . . . . 71
4.3.6 Moderator Explanations or Summaries . . . . . 72
4.4 System Implementation . . . . . 72
4.4.1 Flow A . . . . . 72
4.4.2 Flow B . . . . . 72

**5 Field Study and Demos** **77**
5.1 Feedback from Demos to Harassment Recipients . . . . . 78
5.2 Field Study Methodology . . . . . 78
5.3 Field Study Results . . . . . 79

**6 Discussion** **83**
6.1 Field Study . . . . . 83
6.2 Friendsourced vs. Volunteer vs. Stranger Moderation . . . . . 84
6.3 Harassment on Different Platforms . . . . . 85

**7 Conclusion** **87**
7.1 Limitations and Future Work . . . . . 87
7.2 Conclusion . . . . . 88

**A Survey and Interview Questions** **89**
A.1 Survey Questions . . . . . 89
A.2 Interview Questions . . . . . 95
A.2.1 Email usage [5 min] . . . . . 95
A.2.2 Harassment [25 min] . . . . . 95
A.2.3 Defining harassment [5 min] . . . . . 96

A.2.4 Overall experience with harassment [8 min] . . . . . 96
A.2.5 Responding to / dealing with harassment [8 min] . . . . . 97
A.2.6 Introduce Squadbox [5 min] . . . . . 98
A.2.7 Squadbox-specific questions [30 min] . . . . . 98

# List of Figures


4-1 Diagram of the flow of messages 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. . . . . 65

4-2 On the left, an owner's view of the information page for their squad. On the right, a moderation page for the moderator. . . . . 69

4-3 Squadbox generates whitelist suggestions from an owner's Gmail contacts 74

4-4 An example of Gmail filters for Squadbox . . . . . 75

5-1 Comparison of agreement with statements before and after the field study, where 1=Strongly Disagree, 5=Strongly Agree. . . . . 79

# List of Tables

3.1 Interviewees labeled and grouped based on the nature of 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). . . . . 25

5.1 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. . . . . 79

# Chapter 1

# Introduction

The internet has made remote communication frictionless. We can interact from afar with both people we know and strangers on a variety of platforms, and widely share our thoughts and opinions. We can easily use Google to find information on just about anything, or anyone. We can choose to remain anonymous to those with whom we communicate. While these powerful capabilities have in many ways been positive, they have also empowered bullies and harassers to go after others like never before. According to recent reports by Data & Society [21] and the Pew Research Center [7], nearly half of internet users in the United States have experienced some form of online harassment or abuse.

Online harassment has many forms and degrees of severity, including name-calling, use of offensive language or slurs, physical threats, "doxing" (publicly posting online someone's private information, such as address or phone number), sexual harassment, stalking, hacking, "revenge porn", sending unwanted explicit content, impersonation, and spamming. Harassment can have long-lasting, drastic impacts on its recipients. Research shows that common consequences and reactions are emotional distress, changing one's contact information (such as email address or phone number), deleting social media profiles, and self-censorship of future online posts.

Unfortunately, tools for combating harassment have been unable to keep up. Common technical solutions, such as blocking users and word-based filters, are heavy-handed tools that cannot cover many forms of harassment, are easily circumvented by

determined harassers, and often hinder legitimate, desired communication. Even so, platforms have been criticized for their slow implementation of said features [14, 32].

Recently, researchers have tried to use machine learning models to detect harassment [5, 18, 36], but follow-up work has demonstrated how these models can be easily deceived [16], or contain biases [3].

Given the strong evidence that automated tools are ineffective on their own, a better alternative may be to continue engaging humans in the moderation process. However, many platforms have already put reporting processes and human moderators in place to assist users facing harassment [24], but these have been woefully inadequate. The reporting processes are often cumbersome, and the decisions made on reported content are based on extremely narrow definitions of what constitutes harassment. Harassment recipients cannot currently count on platform action to shield them from harassment [33]. This suggests a need for involvement by humans who are not beholden to the platforms.

We conducted semi-structured interviews with eighteen current and former harassment recipients to understand the nature and context of their harassment, their emergent practices for handling it, and to determine how to best build systems to support their existing strategies. Interviewees came from a wide array of roles, from activist to journalist to scientist, and faced harassment on a variety of platforms. We found that our interviewees considered platforms' existing solutions for harassment insufficient, and that they had often turned to the help of friends instead – employing techniques such as giving friends password access to rid their inboxes of harassment, or forwarding unopened emails to friends to moderate.

These existing practices suggest we should design tools that more effectively facilitate *friendsourced moderation* as a technique for combating harassment. This is the motivation behind Squadbox, a tool that employs friends or other trusted individuals to help people experiencing online harassment by moderating messages according to the harassment recipient's preferences and shielding them from unwanted contact. We use the following vocabulary when describing Squadbox:

- **Moderation**: The process of assessing a message and determining whether

it meets particular standards, or has certain characteristics. For example, "Is this message a threat?" or, "Does this message contain slurs?" In the context of Squadbox, moderation involves deciding whether a message is harassment according to the intended recipient's personal criteria, and deciding which tags from a provided set are appropriate labels for a message.

- **Moderator**: A member of a squad who performs the moderation task described above.
- **Squad**: One or more people serving as personal moderators for another person.
- **Owner**: A user, presumably a recipient of online harassment, who is having their emails moderated by a squad.

Emails an owner receives are automatically forwarded to Squadbox, and then pass through the moderation pipeline. There, the moderator makes an assessment, and then the message is handled according to the owner's specified settings for messages with that particular assessment. For example, some owners might choose to have rejected messages sent to them with a tag, while others might choose to have them vanish silently.

Squadbox has two components: a mail server and a web server. The mail server handles the work of sending and receiving messages, while the web server hosts the Squadbox website, where messages going through the pipeline are stored, where squad owners manage their squad preferences and membership, and where squad moderators do the actual work of moderation.

Given the myriad of ways moderation pipelines are implemented across different platforms, we solicited input from our interview subjects on what features and capabilities would be important to them in such a tool. Perhaps the largest takeaway was that every case of online harassment is very different, and no one particular solution will work for everyone. We spoke with journalists and an academic who were harassed for their work; YouTube personalities harassed for their videos on gender and sexuality; a game designer harassed by a crazed fan; a scientist harassed by their

ex-fiancee; a student harassed for sharing their opinions online; and a victim of SMS spoofing and impersonation.

Some of these users wanted friends to moderate, others paid workers. Some of them wanted to have access to harassing messages; others did not. They all had different opinions on what the most effective way to filter potentially harassing messages would be.

This led us to embrace a philosophy that one of our first interviewees suggested, and later interviewees reaffirmed: "everything should be an option". Rather than making decisions for users about how exactly to use the system, we aimed to make Squadbox as customizable as possible by designing features as options and making the platform agnostic to different possible owner-moderator relationships and usage patterns.

After implementing the system, we demonstrated the tool to five harassment recipients, and received very positive feedback on its current direction. Finally, to explore the privacy implications and personalization afforded to users employing friendsourced moderation for their personal email, we conducted a field study with 5 pairs of friends who used Squadbox for four days. We found that the use of friends as moderators simplified issues surrounding privacy and personalization. However, it also made other issues related to friendship maintenance more apparent, such as the need to ensure moderators feel adequately supported and encouraged in their role by owners.

This thesis details the design and implementation of Squadbox, along with findings from the user interviews and field study. Chapter 2 summarizes related work; Chapter 3 details the findings from the user interviews and their influence on Squadbox's design; Chapter 4 covers the technical implementation of Squadbox. Chapter 5 details the user interviews, and Chapter 6 concludes the thesis and explains both planned and potential future work.

# Chapter 2

# Related Work

## 2.1 Online Harassment Research

There has been a great deal of work characterizing online harassment as a significant problem affecting large portions of internet users [7, 21], with certain groups such as young adults [30, 35], women [30, 9, 28, 31], and those who identify as LGBTQ [21] 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 [21].

As a result of harassment, many recipients simply withdraw from public online spaces [9, 31] or self-censor their content online [21]. Researchers and internet activists have studied or called for better processes to deal with harassment on various platforms [14, 24, 26]. Other researchers examine government policy on online harassment, finding it ineffective [22]. Researchers have also suggested design interventions for platforms to undertake, resulting from content analysis [27], interviews and surveys [31], and design sessions [2] with harassment recipients.

## 2.2 Technical Solutions for Combating Harassment

Researchers and platforms have built technical solutions to combat harassment, beginning with address block lists 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 tried to build classifiers to detect harassing, trolling, or otherwise toxic content, using hand-labeled data [36, 25, 10] or content from existing communities [5]. Researchers have also worked to release data [12] and to better define subtasks within the overall space [18, 34]. 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 [16]. 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. However, these approaches may still have a role to play. We build upon this prior work by considering how automated approaches could augment the work of human moderators.

## 2.3 Community-Based Systems for Combating Harassment

By building on prior research methods and findings [8, 23], 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 [11]. Some of these tools

include BlockTogether [15] and Good Game Auto Blocker [13], where users collaborate on blocklists of harassing Twitter accounts and can automatically block profiles based on attributes such as the age of the account. Other community-based efforts include projects such as Hollaback! that elevate victims' stories [6], 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]. Prior work on HeartMob finds that having an external party classify messages as harassing can be a validating experience, and that participating in this classification process motivates community members to help further. 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 studying harassment recipients' existing strategies in order to design a tool that augments those strategies.

## 2.4 Collaborative Systems for Message Management

Finally, we draw from research on systems for collaboratively managing and moderating messages to and from individuals. Our research group explored email usage in mailing lists, finding use cases for friendsourced moderation of one's outgoing email to overcome anxieties about posting messages to large public lists [37]. 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 [19, 20], 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 the context in which messages are sent and received. Privacy considerations for friends are also significantly different than those for strangers, which we explore in our user interviews.

# Chapter 3

# User Interviews

We began 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. This chapter describes our methodology and interview subjects, followed by our findings and their implications for designing anti-harassment tools.

## 3.1 Methodology

We conducted semi-structured interviews with eighteen potential users throughout the design and building process. Through social media and our various professional networks, as well as by cold-emailing people featured in news articles and blog posts, we sought out people who (according to their own definition of "harassment") had experienced online harassment on social media or other communication channels to participate in 45-minute to 1 hour long interviews via Skype, phone, or in-person. The interviews were conducted by the author along with another graduate student working on the project. In addition to the interview, participants filled out a brief Google Forms survey to gather demographic information and some basic data about their internet usage and experience with harassment.

We developed a set of questions ahead of time to guide the conversations with our subjects. The first half of each interview focused on understanding interviewees' experience with online harassment: the who, where, and how, as well as the impacts

harassment had on their life and actions they had taken in response to it. In the second half, we turned to discussing how interviewees might use a tool for friendsourced moderation to combat harassment. The full list of interview questions, as well as the questions contained in the survey, can be found in Appendix A.

We transcribed interviews and performed a qualitative analysis of the interview transcripts, using a grounded theory-based approach to code the data and develop common themes. Some details and quotes have been changed or edited to protect the identities of our interviewees. We use "they" and "their" as singular pronouns to protect the gender identities of individuals.

## 3.2 Interview Subjects

Sixteen of eighteen participants completed our survey to gather demographic information. They ranged in age from 18 to 52, with an average age of 33.25 and a median age of 31. Eleven subjects identified as female, two as male, and the remaining three as non-binary/genderqueer. Twelve subjects identified as white, three as Asian, and two as Middle Eastern/North African. They all resided in the United States. Their experiences with harassment varied widely, as shown in Table 3.1. In that table, we show their occupations and the main platforms they experienced harassment on. We also give a brief description of the nature of their harassment, along with peak harassment volume and average harassment volume (in terms of number of messages/comments received.) The provided labels grouping participants into high-level groups based on the nature and trigger of their harassment are used to tag the quotations in this section.

## 3.3 Findings

In the following section, we describe the common themes we found in interviews, grouped into the following five categories: 1) how interviewees defined harassment, 2) how harassment affected interviewees, 3) actions interviewees took in response to

| 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 politics/social issues 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 follow 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 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 & FB because of activism on controversial & identity-related topics, & 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 video content. 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 video content. Has been doxed. | 50+ | 50+/day on YouTube and Twitter |
| YouTube personality [You3] | Email, Twitter, YouTube | Identity-based attacks and threats based on video content. Has been doxed. | 50+ | 10+/day on YouTube and Twitter |
| YouTube personality [You4] | Facebook, Instagram, Twitter, YouTube | Identity-based attacks and threats based on video content. 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 w/ dissenting political opinions for views in 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 3.1: Interviewees labeled and grouped based on the nature of 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).

harassment, 4) what interviewees wanted to change on existing platforms, 5) how interviewees thought other people could help them, and 6) what features and options interviewees wanted in an anti-harassment tool like Squadbox.

### 3.3.1 How Subjects Defined Harassment

We asked: "what sort of things do you think constitute harassment?" While individual experiences varied greatly, the two features that came up most frequently were 1) insulting and/or threatening content, and 2) repeated, persistent contact.

#### Message Content

In terms of message content, two commonalities emerged in interviewees' definitions of harassment. The first was that harassing messages are personal attacks, often about the recipient's identity. Interviewees said that when these attacks occurred in the context of discussions, they were usually irrelevant to the topic at hand. The second commonality was that the message's intent was to be upsetting. Regarding identity-based attacks, interviewees said harassment was:

> **You1**: *Repeated attacks that are based on immutable characteristics, like gender. People make fun of my weight a lot. People make fun of how I talk. Things that I can't really control. These are all things I would consider harassment.*

> **You2**: *Attacks on someone's identity, and cultural history. I'm [religious affiliation], and that's been a big thing that people talk about and get me for.*

> **Jour1**: *Anywhere where the point of the message veers into a personal attack.*

And regarding the sender's intentions, they saw harassing intent as:

> **Act1**: *If someone is contacting you with the intent to upset you and disturb you, I consider that harassment.*

#### Message Patterns

Outside of the message content itself, subjects noted that volume and frequency of messages were a large factor in their feeling they were harassed, and that even

innocuous-seeming messages could be troublesome when arriving in large volumes. Interviewees stated:

> **Res1**: *Someone keeps messaging you, and you clearly – through lack of replies, or through very short modest replies – don't want to engage with them. If someone just sends me one message... I wouldn't call that harassment. But if someone keeps sending me things... I would say this is harassment.*

> **Act1**: *If you tell someone not to contact you, and they continue to contact you, that is harassment.*

> **You3**: *It's repeated things, trying to get a negative response.*

You4 echoed that sentiment, noting the persistent nature of their harassers:

> **You4**: *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.*

### 3.3.2 Consequences of Harassment

We asked interviewees how their experiences being harassed had affected their lives and their use of the internet. The consequences they described largely fell under the following two categories: 1) detrimental effects on mental health, and 2) alienation from online communities and conversations.

#### Effects on Mental Health

Overwhelmingly, interviewees said that they suffered emotional and mental consequences from reading the contents of harassing messages. Anxiety and depression were common results; multiple interviewees reported that they began seeing a therapist because of their harassment.

Describing the peak of their harassment experience, Act1 told us about the effects at the time and the lasting consequences:

> **Act1**: *I always had underlying anxiety issues, now I have huge anxiety issues. I now take antidepressants and have a therapist... all of that was triggered by harassment. I was shaking every day. I was so upset I couldn't go online. I just felt horrible about myself and scared. I spent months in a severe anxiety*

> *mess. Not going out, I stopped writing, I took a break from activism. I was just really struggling, and even now, I just feel so depleted. It's just so exhausting. It's definitely had such a huge impact on my mental health... I'm so worn down emotionally and psychologically.*

Another interviewee said:

> **You2**: *It's taken me a long time to figure out how to deal with this...I go to therapy, two sessions a week. That's a very important thing for my life, but I would have not gotten through this past year with all this harassment if I didn't have a space where I was prioritizing my mental health.*

Echoing the sentiment about being worn down, You4 told us about how the constant nature of online harassment was so difficult to deal with:

> **You4**: *It's been shitty for my mental health... The constant negativity really got to me for a while. Just having it in your mind every 30 minutes or whenever there's a new message. You can block them, but just being reminded every few minutes, "Hey, you're horrible," really got to me for a while and I just felt really really bad and didn't want to make YouTube videos, didn't want to tweet or anything because I knew it would just make it worse. it just wears me down.*

### Self-Censorship, Hindered Communication, & Alienation

We asked: "has harassment changed the way you communicate online, or the way you use the internet in general?" The common threads that emerged from this were: 1) interviewees self-censored and shared less than they wanted to in order to avoid future harassment, 2) interviewees thought harassment severely limited their ability to participate in online communities, and 3) harassment made it difficult to have personal communications.

**Self-Censorship** Several interviewees stated that harassment had indeed changed the way they communicated online. It led them to think more carefully about what they posted, and to avoid making some posts altogether:

> **Act1**: *I think very intentionally about everything that I post. Which accounts are public and which are private. I censor myself so much, in terms of I can't say anything too outlandish, because if it gets taken out of context and put in an article, then that'll be a whole issue... It's definitely limited the way that I interact publicly.*

In particular, some interviewees felt that harassment had hindered their ability to talk openly online about their personal lives to the extent that they wished to.

> **You3**: *It's definitely made me be more conscious of how I'm using social media platforms... It made me be more aware of not sharing as much as I would want to, because if I had shared more personal stuff online that's going to open up more harassment.*

> **Act1**: *I do censor myself a lot. I can't talk about a lot of things that are going on in my life that I wish I could. I wish I could tweet more openly about discovering my sexuality, and struggling with mental health, and my parents getting divorced. There are a lot of things I wish I could talk about, but I don't anymore because I know it will be used against me and it could hurt the people in my life.*

Res1 noted that they'd begun to blame themself, rather than the harassers, when something they posted led to their being harassed:

> **Res1**: *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.*

**Alienation from Communities** Outside of having their own posts hindered, many of our interviewees said that harassment required them to make themselves harder to contact by closing their direct messages (private messages on Twitter, which the user can choose to disable for all but people they follow), turning off notifications, disabling comments, etc. They said that while this helps mitigate harassing contact, it also makes it far more difficult for them to engage with the people they do want to talk to – both people they already know as well as non-harassing strangers.

In particular, the YouTube users we spoke with felt they were unable to effectively communicate with their fans and fellow YouTubers:

> **You1**: *It has been very disruptive to my work in terms of my ability to teach and to cultivate a positive space for my viewers. It's made it really, really hard to do that and almost impossible without me having to subject myself to a lot of vitriol every day.*

> **You2**: *I'm less inclined to engage. I would love to really build a strong base of fans and a community, and that's really hard to do because I have to weed through all this crap.*

> **You4**: *I've had to shut down a bunch of avenues for reaching me. My Twitter DMs used to be open, and I used to reply to people on Instagram and Tumblr, and now none of that happens. I talk to strangers a lot less just because having those doors be open is too much, so I just have to shut them completely. It's either be constantly bombarded with harassment all day, or maybe find a new friend.*

You3 even noted that some of their fans were afraid to publicly interact with them for fear they'd receive harassment by association:

> **You3**: *It's made it really difficult to interact with the people who don't want to harass me. People message me and tell me that they've been afraid to comment or interact with me on platforms that are public, because they're afraid of also getting harassment. 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 check my Twitter notifications from people I don't follow, or go through my emails. Why should I look through hundreds of harassing comments to find a few good ones?*

On the other hand, Res2 states that while facing online harassment was traumatic, it led to an outpouring of support from their community, led them to meet new people, and ultimately helped their academic career:

> **Res2**: *I don't want to downplay how bad it is. But this was actually really good for my career. The fact that my university really stood up for me, the fact that people in our local community really stood up for me, people in my association... Within a year, I was on a keynote panel next to the other really renowned people in my field. I guess I want to balance that with, it's important that people understand how serious it is. It was extremely stressful for me at the time, very stressful and very scary... But for me so many people stood up for me in so many different ways in my immediate circle and through the university. I was a front page story in my local newspaper, the mayor wrote to me... I had, at the national level and international level, many people that I respected who I had never been in touch with writing to me and saying "hang in there, don't worry." So those things really made a big difference.*

**Hindered Communication** Some interviewees expressed that harassing messages made it harder for them to communicate with people they already knew. This was often because they were overwhelmed by a high volume of messages, effectively experiencing a Denial of Service (DoS) attack [21]. Oftentimes, this harassment was incited by a particular person with a large following, who could direct "hate mobs" to flood their communication. One such case was Res2 – they told us how someone

else at their university had been tasked with deleting the harassing messages during the peak of harassment. Even so, the messages made it impossible for Res2 to use their email account:

> **Res2**: *I couldn't check my email. I wasn't even checking my email really during that time because there were so many coming, even though the person was [deleting messages] several times an hour.*

On the other hand, while Spoof1 did not experience a high volume of messages, harassment still impeded their communication because they became unable to distinguish between legitimate messages from friends and spoofed messages from their harasser(s).

Interviewees worried that they were missing out or could in the future miss out on opportunities as a result of harassment. One noted that online harassment led them to make themself difficult to contact:

> **Act1**: *I think that there are a lot of opportunities that I miss because it's impossible to contact me if you don't have my contact info. It's really frustrating that, because I'm worried about getting harassment, I can't be available to journalists as a source. I used to get journalists being like, "Hey, can I get a quote from you on this?" All the time... I used to get all these awesome opportunities and I just can't get them anymore because people will use my email address to send me terrible things.*

Another interviewee, Jour3, noted that they had in the past missed an interview request via tweet because the volume of harassing tweets was so large.

### 3.3.3 Actions Taken in Response to Harassment

We asked interviewees about what actions, if any, they had taken in response to previous harassment. In the following subsections we detail four categories of responses: 1) blocking and filtering out harassment, 2) responding to harassers, 3) reporting harassers, and 4) asking friends and family for help.

#### Blocking and Filtering Out Harassment

Nearly every subject we interviewed stated that they had blocked social media users and/or specific email addresses in response to harassment, though most subjects felt

that this mechanism was not very effective, either because they had many different individuals harassing them, or because determined harassers can easily circumvent blocking by making new accounts and sometimes new identities:

> **Act1**: *Most people aren't repeat offenders. Most people just attack you once. And if someone really wants to get at you, they can just logout of Twitter or go in incognito mode or open a different browser and continue to bother you and read what you're doing.... It's very easy to create new Twitter accounts. It's very easy to create new email accounts. If somebody gets blocked, sometimes they'll get even more pissed. It doesn't necessarily help at all, really.*

> **Res1**: *There's two people who keep creating new accounts, and they keep messaging me and ten of my friends. They're really obnoxious. I've never met them in real life... And every time they create a new account, I just immediately block it. I can tell pretty quickly.*

> **You3**: *They're definitely people who are committed to constantly harassing me and other people. I'm surprised in that I block so many people... and then there's more people who keep commenting for the first time. Most of it is not repeated on a longer basis because I notice it's happening then I block them.*

> **Fan1**: *I blocked that email. How he got back in contact with me is creating a new email and then contacting me through that. I finally blocked his new email, and I didn't hear from him for a few more months. He did the same thing again, and created new emails to try to contact me again, and I just blocked every single new email that I got from him. Every time he makes a new email, he creates a new name as well, so I can't block a single name, I can't block a single email... Not only new names, but he also pretended to be different people. He keeps winning my trust, again and again, and then revealing he was the guy from before.*

A few subjects expressed hesitation to block harassers, feeling it was important that they read or at least be aware of their harassing messages. One reason for that was a desire to be aware of incoming threats:

> **Jour2**: *I've blocked some of these people who have threatened me, but I always worry that because I block them I'm gonna miss something I should know about... Just in case they are threatening me, I need to be aware of it.*

Another reason we heard from journalists was that they felt reading dissenting opinions, even if written in a harassing manner, was crucial to doing their job well:

> **Jour1**: *Tough criticism is okay. Even if there's a few profanities or whatever. I think it's my job as a journalist to be able to answer that and assure people that I want to hear their side, that I want to be fair.*

> **Jour2**: *I read them all still... even the ugly ones are helpful to me in my work. I read them specifically because I need to know what people are thinking to be good at my job. A lot of people complain that reporters or columnists are in this bubble. I'm not in a bubble. I always have a very clear picture of how people who disagree with me are thinking and the issues that are important to them.*

Interviewees also noted that word or phrase-based filters had been inadequate. Some expressed frustration at being unable to come up with the right words or having the filter catch innocuous messages that simply quote a harassing message or use the word in a non-harassing way, such as towards reclaiming slurs. One subject described this as:

> **Act1**: *The only problem is sometimes people will use those words thoughtfully. I've had friends use those words, too, and with all of the moves to reclaim slurs... Especially things like slut and whore and cunt*

One interviewee described filtering out messages with some words despite knowing the filter would catch non-harassing messages, saying

> **You3**: *I also 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 as a topic of conversation. If I have the energy to, I'll go through my "held for review" folder to look through those...*

### Responding to Harassers

Some subjects had tried responding to their harassing messages in the past. Their opinions on the success of this were mixed. Some felt that it could have a positive impact and stop the harassment:

> **You1**: *I have had some success in talking to people who tweet me those kinds of things. And they usually seem to have an a-ha moment, as if they didn't realize that there was another person on the other end of the line.*

Others had instead come to feel that it was just a waste of their time:

> **Res1**: *I used to respond to all of them all the time, and write really long things. Sort of over the past six months to a year, I don't. It's just so much time and energy that if this kind of thing happens to me now, I will just not respond.*

### Reporting Harassers

Nearly every subject had reported or attempted to report harassers to communication platforms; a large number had also sought help from outside sources, such as law enforcement. For both platforms and law enforcement, subjects strongly expressed disappointment and frustration at the reporting process and the platforms' responses.

A common frustration with Twitter's reporting pipeline in particular was that reporting is an inefficient process with a lot of work involved:

> **You2**: *Their method for blocking and reporting is really intense whenever you go through it individually. They send you all these emails back... They ask you to send them an email back detailing it... Maybe I should've taken screenshots, but I also don't want to have that on my iCloud. There's just a lot of burden on the harassed person to take care of it themselves.*

> **You3**: *You report it and they email you and then you have to provide evidence of why is this harassment, which is just a cumbersome process... it's not worthwhile to go through that each time someone is harassing me.*

Users also expressed confusion at how platforms made decisions regarding what content was allowed. One particular user gave up on reporting altogether as a result:

> **Res1**: *I definitely have reported people before. I did it more often than not two years ago. And then once I realized that things actually didn't happen... I clearly don't know how to use this, so I won't.*

While social media platforms at least have (problematic) reporting mechanisms in place, one subject tried to report harassment on her Gmail account to Google and couldn't even figure out how to do so:

> **Act1**: *I don't really know if there's a way to report to Google. It's a really complicated process... It was very strange and I had to really dig to find the mechanism, so I haven't really done much... I don't really know where to start.*

Among those seeking help from law enforcement, the prevailing sentiment among them was that it was an unfruitful experience. They cited lack of interest and action from the police as their primary frustrations:

> **You2**: *I've had one bad death threat and I called the police and they just didn't really do anything. It just didn't seem like they could.*

> **Res2**: *I was worried for my personal safety because of some of the emails I received. The police didn't really have anything they could do about it. They didn't seem interested.*

> **You4**: *I have had some with people who are just like, 'I'm going to kill you. I'm going to find you and I'm going to kill you.' That should be a thing that I should be able to go to the police with. But they just don't care because it's an internet thing.*

Subjects felt that law enforcement was incapable of and uninterested in taking any real action in response, and that their concerns were not taken seriously because of a disconnect between "real life" and the internet. As was the case with reporting to social media platforms, a lack of response from police discouraged subjects from reaching out or reporting again:

> **Act1**: *I never heard from the police ever again after filing my report. I don't think they ever investigated it. I've gotten a lot of emails since then, so I should update the police report, but it was such a shitty experience that I just haven't had the willpower to go back.*

### Asking Friends or Family For Help

We asked interviewees if they had ever asked people they knew – friends, family, coworkers, etc. – to assist them in dealing with harassment in the past. The majority of them had. The help they received from other people largely came either in the form of emotional support, or in a more tangible form that involved actually dealing with the harassing messages.

**Emotional Support** Interviewees felt on the whole that their friends had been very supportive of and helpful to them as they experienced harassment – as much as they might be in any other emotionally difficult situation. Fan1 said:

> **Fan1**: *I felt like my friends were really, really helpful... I think it was my friends that finally got me to realize that this was definitely abusive.*

Act1's friends helped by becoming advocates on their behalf, and defending them when people said they were overreacting:

> **Act1**: *Some of the best ways my friends and family have helped me is just by calling other people out when they say things like that, and explaining to them, 'Oh, no, no. You need to understand what this is like.' This can be incredibly traumatizing. So, being an advocate and breaking down those misconceptions.*

**Moderation-Like Support** Several interview subjects described ways in which friends had digitally helped them manage harassment by literally or at least effectively moderating their messages.

Act1 said that their best friend had their Twitter and Facebook passwords, and when Act1 was going through a particularly bad bout of harassment the friend logged in to Act1's accounts and cleared out harassing messages and notifications, and blocked users. Act1 said the same friend also officially served as a moderator on Act1's public Facebook page.

Res2 said their spouse would log in to their email account and delete the harassing messages. You4 said that their significant other would go through the comments on their posts, which were largely negative, and read aloud to You4 the more positive and encouraging ones.

Multiple interviewees said that they would have friends read potentially harassing emails first, by forwarding them unopened:

> **Act1**: *If I see an email come in from an account that looks like it's from my stalker, I'll forward it to a friend and be like, "Can you read this for me and tell me if there's anything in here that's new?"*

### 3.3.4 Desired Platform Changes and Features

We asked interviewees: 'How do you think platforms could help you deal with harassment? Are there changes to the interface or additional tools that you think could help?' The main changes interviewees desired were: 1) help with "hate mobs", 2) better harassment detection features, 3) improved reporting processes, 4) more control over content and comments, 5) more customizable notifications, and 6) more thorough blocking mechanisms.

### Help with Hate Mobs

Interviewees who had experienced harassment in more public forums like Twitter and YouTube told us they wished platforms worked harder to fight against 'hate mobs' – essentially large groups of platform users simultaneously going after another user. They told us these were often fueled by someone with a large following directing their followers toward that user:

> **Act1**: *I think the issue is that it's very easy for someone to create a hate mob and send their followers after someone. It's very difficult to create a designed solution for that.*

> **You3**: *I think a big thing with harassment is the inciting harassment thing. Platforms need to realize that's a real thing and target the people who are inciting harassment. There are millions of people who use social media to harass others, but there's a smaller number of people who use social media to incite harassment, so it's more impactful to target the people who are inciting harassment.*

> **You4**: *Instagram, Facebook, Twitter, all of these social media platforms need some way of identifying people with large followings who are using that following to attack people and ban them.*

In some very public cases, platforms have responded. For example, the alt-right commentator Milo Yiannopoulos was permanently banned from Twitter after urging his followers to harass actor Leslie Jones after the new Ghostbusters movie came out. But less famous recipients, such as our interviewees, have much less success with getting the inciters of harassment banned.

### Improved Reporting Processes

Many interviewees were frustrated with the processes for reporting harassment to platforms. They saw this as a major place platforms could improve when it came to helping them deal with harassment.

Act1 said, regarding email specifically:

> **Act1**: *Having a more transparent reporting process, having more accessible language when it comes to, "How do you report a stalker on Gmail?" These are not uncommon problems. I'm sure they happen all the time and just having more*

*accessible human FAQ pages, really anything. Having email addresses that are transparent and that people can contact ... I just know that the magnitude of messages they would receive is a lot, and it takes a lot of investment on the part of companies to build those teams. But I think that it's worth it when it comes to the safety of your users.*

You3 and You2 called reporting harassment on Twitter "a cumbersome process" that places "a lot of burden on the harassed person".

### More Control Over Their Content

Something we heard from all of the YouTube personalities we spoke with was that they experienced harassment via 'response videos', videos that other YouTube users made in reply to their videos. These videos contained attacks on the original poster. Since the videos are considered 'similar' according to YouTube's algorithms, they show up in the sidebar as 'related videos' for users. Interviewees expressed that this was negative for both them and their viewers:

**You1**: *What happens is I'll post a video about sexual assault or something, and then I have a bunch of people, men's rights Act1 types, who'll make a video saying that sexual assault isn't a problem, etc. And they are often quite vitriolic, call me names, create degrading cartoons of me. And those will all be on the sidebar, next to my video.*

**You4**: *The sidebar of all of my videos are just horrible, horrible things. I get messages from people all the time that are like, "I like watching your videos, but as soon as I watch one, the next 10 recommended videos that I get are like 'why You4 is a stupid idiot,' etc. It ruins my viewers' experience, and I can't do anything about it. And every time I go to watch one of my own videos, these things are popping up on the side.*

**You2**: *I was really worried that whoever was watching would click on them thinking it was my videos, and then people are just cursing and saying all this mean crap, and that's pretty traumatizing for anyone who's coming from this nice show to this horrible bullying thing.*

Users cannot control what shows up as related to their videos, and our interviewees expressed a desire to block particular channels from showing up there. Users can disable the 'related videos' section completely, but You2 described the issue with that:

> **You2**: *When I turned [the sidebar] off, my videos can no longer be suggested in someone else's sidebar, so that works against me now. I'm being penalized for protecting myself and my viewers.*

Similarly, channel pages display recommended similar channels. You4 said this section would often include the channels that made response videos to them:

> **You4**: *You can't choose who it recommends. Mine is currently recommending horrible people who make violent videos about me. You can't add or delete or remove them. If you turn off that feature – you can just turn it off completely – then you don't get recommended on anyone [else's channel page], so it hurts your channel's performance, and you're not able to get recommended even on good channels.*

And just as with the sidebar feature, users could not disable this feature without hurting their own traffic.

You4 felt strongly that, if they were the one posting a video, they should be able to determine what was shown on the same page as it, just like on other platforms:

> **You4**: *Youtube is, at its core, a social media platform. You should be able to control what's on your page. Facebook doesn't recommend other shitty Facebook pages next to you, and if they did, you could click the little X and be like, "Don't show this anymore." But you have no control over that on YouTube.*

### More Customizable Notifications

Interviewees stated that while some platforms allowed complex customization for notifications, other platforms could stand to improve.

> **You3**: *Twitter has done better because they have a way to set notifications for only people you follow. That would be a great thing on other platforms, to be only people in your contacts or only people in your circles, if it's Google Plus or whatever. That's a good feature.*

### More Thorough Blocking Mechanisms

Interviewees told us it was difficult to truly block people from contacting them again. Although many platforms allowed users to block accounts, that was easily circumvented by motivated users who created new accounts to continue targeting them. Additionally, You4 told us that YouTube's block feature lacked several important

traditional components of blocking. Blocked users can still view your content, subscribe to you, and like or dislike videos:

> **You4**: *YouTube's block feature is just non-existent. You can block people on YouTube, but you also have to block them on Google Plus, otherwise they can continue to comment on the video on Google Plus... The process to do that takes like 500 clicks, and you can only do it on the desktop. There's no way to do it on mobile, so it's time consuming and slow... The only thing it does is stop people from commenting. They can continue to be subscribed to you, they can continue to dislike your videos, all of this stuff. Several of the hate channels that make videos about me are still subscribed to me. I can see that they're subscribed to me. The second I upload a video, they get a notification and they can make shit about it instantly. Like Twitter, you block someone, they can't even look at your profile. Why can people still watch my videos?*

### 3.3.5 Desired Assistance from Other People

Before introducing Squadbox, we asked interviewees: 'If there was a group of people that were willing to help you deal with harassment, what do you think they could do? What areas would you like assistance with?' Of interviewees that had ideas of others could help all of them explicitly stated, or at least described moderation:

> **Act1**: *I think that being a moderator, being like that first ball, is incredibly helpful. Having people read Facebook comments, remove things, block things. I've made use of that. 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.*

> **Jour2**: *I would love to have another name, but without feeling, read all of the emails, and sort the useful stuff from them, and take the other stuff away from me.*

> **You2**: *Having moderators would be great. People who I can trust that who I know are secure enough to be okay with it. 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.*

> **You4**: *I guess the only thing that would really need to be done is sorting messages that I should see and messages that I shouldn't see. Things that are useless and attacking and harassment and just sorting those out, and things that are questionable or positive or questions or whatever, then that I can see. That would be great.*

You1 said that, although moderation would be helpful, they had been unable to find someone to moderate in the past:

> **You1**: *A human could moderate the comments. I did try to hire someone at one point. I talked to probably four or five people and once they realized what I was asking them to do, nobody wanted to do it, which I do not blame them. Maybe because it was my friends, people who are in the digital sphere who do social media management, it's just too much toxicity. But figuring out how to have a human moderate would be helpful, for sure.*

### 3.3.6 Friendsourced Moderation Feedback

In the second half of the interview, we turned to discussing our idea for a friendsourced moderation tool, and solicited feedback and suggestions on its design and proposed features. The description of Squadbox we gave to interviewees is as follows:

Squadbox is a tool to combat online harassment via email. It places a "squad" of moderators in between you and the people who email you. Your squad is made up of one or more moderators, who review the messages sent to your Squadbox and decide whether the messages should be approved, rejected, or tagged, according to your preferences. The messages you want to see are then sent on to your inbox.

We summarize the interviewees' feedback in this subsection; in the following chapter we discuss how how that feedback led to our final design choices.

#### Who Moderates

We imagined three different groups of people could serve as Squadbox moderators - friends/family (unpaid), paid strangers, and volunteer strangers. We proposed these ideas to our interviewees and asked their thoughts on the different groups, as well as which they would be more likely to use, and what privacy concerns they would have with different types of moderators. Their answers varied widely.

Some interviewees expressed that they had privacy concerns when it came to having moderators they didn't already know:

> **Jour2**: *I get a lot of personal emails through my work account. So, I guess I would prefer it to be people I know.*

> **Act1**: *[Strangers is] where I would be more uncomfortable, actually. I feel like getting harassed is such an emotionally fraught experience sometimes that I prefer to turn to friends for support. It can be very violating, some of the emails that I get, and it almost feels more violating to have somebody who doesn't know me read those somehow. That's where I would worry about personal information. You never really know what you don't want someone else to see until they see it. I probably wouldn't use a service like that.*

Three of the YouTube personalities we spoke with stated that they would want paid moderators, as their harassment was directly related to their jobs and source of income. They expressed that they would feel guilty placing that burden on friends:

> **You4**: *It might work for some people. I would feel weird about it. I would feel weird paying them, and I would also feel weird not paying them. It's like they deserve to get paid to put in that labor, but they're my friends, so it's weird to pay them. Then I don't want them to volunteer and do it. Then it's like they're doing work for me. I'm sure there are people that would work for. I am probably not one of them. Strangers I could get on board with if I had the budget to pay people to look through it, pay strangers to look through it. Yeah, I could get behind that. That'd be really useful.*

> **You1**: *Because this is part of my business, paying people to do it is the most appealing to me. And having a way to find those people very easily without having to search myself would make it really ideal. Because this is my business, I don't want to ask my friends to basically do work for me, or a volunteer to do work for me. This is my full time job, so just as a business ethics thing, I would use the paid moderators.*

> **You3**: *I feel bad asking people to volunteer their time, or even asking family to ... I don't want them to know the stuff that I go through because I don't need to see that, because I know it wouldn't be healthy for them either. I think the ability to have comment moderators on YouTube is such a useful feature that every social media platform should really have. It's just something I haven't used on other platforms, because I would wish I could pay people to do that for me so they're not just volunteering their time to look through hate. I prefer it to be a more business relationship than a friendship support type thing.*

The fourth YouTuber, You2, stated that since their email harassment volume was rather low, they wouldn't be as concerned about asking a friend as the others. However, they noted that a paid, non-friend moderator might be preferable, as they were choosing to do the work rather than being asked to do it:

> **You2**: *The volume is a lot less in my email. I don't know what would be different if my volume was really high. I think it's the only thing is that moderating emails is a lot more time consuming than a YouTube comment. I guess that would be my only concern in terms of burdening someone else with this. I think strangers would be great. It's a source of income for someone, they're essentially volunteering to do it rather than being asked. That's a way for someone to pay privilege forward if that's something they're interested in, they're specifically signing on to have to read all of this crap. I think there's a big difference between being asked to do something and volunteering yourself. And if you're going to pay for it, even better.*

Regarding privacy concerns with non-friend moderators, the YouTubers had some, but felt they could be mitigated through transparency to people contacting them and by getting to speak with and know their moderator:

> **You1**: *There could be a privacy issue because people sometimes email me private stuff thinking it's going just to me. But you know, I could just put a note about that on the [contact] form [on my website] to let people know that it's going to a team and not necessarily directly to me.*

> **You2**: *I think the stranger would need to be vetted. I would want to talk to them personally, just to make sure I can trust them on a basic level and make sure we're on the same page about what, for me, constitutes harassment and what is okay to send my way and what is not. And making sure that they believe in what I'm doing and don't want me to be harassed for it.*

> **You3**: *I feel like if I were to hire someone to moderate emails for me I would want to make sure I trust a person and get to know them a little bit first, to not just hire some random person.*

> **You4**: *I don't have any, but I'm also just not ... I never have privacy concerns. I let Microsoft know all of my location data and targeted ads. I don't care. That's just me. I know a lot of other people do, but no, I don't care at all.*

Despite their interest in paid moderators, it is important to note that two of the YouTube personalities, You3 and You4, stated this would not be financially possible for them.

### Moderators Communicating with Harassers

We asked interviewees whether they would ever want their moderators to communicate with the people sending them harassing messages. Their answers were mixed.

One suggested that perhaps being told to stop by someone else could be more impactful:

> **Res1**: *it seems kind of like a very interpersonal thing, just in terms of if someone is bothering you in any way, and you can say whatever you want to them, but they probably won't hear it even. But if someone else says that to them, they might hear it. Like, "Hey, stop bothering my friend," sometimes that for whatever reason will be a pseudo wake-up call for the person who's doing it.*

Similarly, another thought moderators could be useful for defusing the situation and reminding the harasser their recipient is human:

> **Jour1**: *I would love it if I could have someone else to communicate with these people to defuse the situation and see if they can get to the heart of the issue. If it's someone that's reasonable, that can be calmed down and find out what their issue really is, I would love some assistance with that... Sometimes it's surprising how someone sends a hateful, harassing email to a journalist and they're always surprised when they get a quick answer that is professional and courteous. Often they'll respond with, "Oh, okay, I appreciate you answering me but I don't agree with you. I still mean what I said but I appreciate your answer." I can see my surrogate using that approach... I think people need to be reminded that journalists are people, we have feelings.*

Act1 thought moderators could help educate their harassers, or send helpful information to strangers, or tell off the harassers:

> **Act1**: *What they would respond to would depend, but I do think it would be really fun if I could have people educate some of the people who message me. Because sometimes it is just someone who's ignorant. Who sends me an insensitive email. I worry if I were getting an email from someone who really wants support but shouldn't expect that of me... It would be really nice to have a moderator to be like, "Hey, Act1's not going to answer you, but here's the resource list that they've prepared. Or like, "Here's some things that we can give and here's a help line." So that I don't have to do the emotional labor of reading those messages, but they're still getting something that can help. Or if somebody's just being really gross and insensitive, write back and be like, "This is Act1's friend. You are appalling for X reason.'*

You1 was unsure, but open to having their moderators try it:

> **You1**: *I don't know, I think I'd want to experiment with that. I have had some success in talking to people who tweet me those kinds of things. And they usually seem to have an ah-ha moment as if they didn't realize that there was another person on the other end of the line.*

On the other hand, some interviewees felt communicating with harassers might be unproductive and actually lead to further harassment:

> **Fan1**: *No. Do not feed the trolls. I trust my friends a lot but also I'm really emotionally bonded with my friends. I also feel like if they had the option to reply to a troll for me, they would do so, giving them ... whaling on them. I get it, my friends care about me, and they want to fight back on my behalf, but that's not what I want when dealing with trolls. Don't feed them.*

> **Res1**: *I'd say just like not engage. I think like ... It's like the older internet phrase, "Don't feed the trolls" kind of thing. I think the less engagement, the less someone on the other end has to actually work with.*

Another interviewee did not want their moderators communicating with people on their behalf, but suggested a pre-written reply could be useful:

> **You2**: *Probably not, unless there's some sort of, I'd say maybe automated response of, you're harassing me, please stop, this is why this is bad, something like that they'd have to send back. Otherwise, I wouldn't, just because I would want to make sure that it was a brand voice, and I would want to control that.*

Spoof1 similarly stated that they did not want moderators communicating on their behalf without their input and consent.

### Automatic Replies

We proposed to interviewees that Squadbox could have a feature where, the first time someone contacts you, they get an automatic reply. This could notifying them that the message was going to undergo moderation, and give them the chance to revise or rescind their message. Interviewees were mixed on this. One thought that it would discourage their harassers and encourage people contacting them to be more polite:

> **Jour2**: *I think it would be really helpful actually, for people to know that the emails are being moderated by somebody... I think it's okay to do the re-phrasing thing. I think it will have the same effect as rejecting the email for some of these people.*

> **Res1**: *Yeah, I think people should always probably reread their messages. I think the option to always have that second is always good.*

The interviewee who was only harassed by their ex agreed, saying:

> **Ex1**: *I think it would be great if [my ex] knew that they were being moderated because it's a sign that it's a step that I've taken to put distance between us. If they know that the things that they're sending me aren't getting through, they're likely to stop sending them.*

Some interviewees were unsure what effect it would have, expressing that it might lead them to just be harassed more on different platforms, and give their harassers the satisfaction of getting a response:

> **You1**: *I feel like that would probably limit the number of emails I got. I feel like there would be a lot of people who would realize that ... Would understand that they're just trolling and then would realize that their email would not get to me and wouldn't send it. I could see that also having backlash... often times if I have turned off comments on a video then I will see more harassment on my Twitter or more harassment on my other videos from people upset that they couldn't comment on that one video. I could see some people upset by it and then using that as motivation to harass on a different platform. I think it would be something I would have to test out to know if the effect would be stronger of people deciding not to send, or that upsetting them and then sending more. I could see it either being useful or counterproductive.*

> **Act1**: *I'm always really curious how people write and act differently once they've been prompted to think about their behavior... I used to have an auto reply set up on my email to be like, "Hey, I'm not going to get to this right away and I don't reply to everyone, but thank you." I found that sometimes people would be even more pissed when I didn't respond because they knew the email went through. I think in some cases it's really awesome and in some cases it can backfire when somebody really wants to contact you.*

> **Res2**: *I certainly felt like they're trying to get a reaction to them. But if it's an automated reply it could be kind of demoralizing and so maybe that would be a good thing. Especially if it looked really automated.*

Others leaned toward the side of thinking it would definitely cause more harassment and just make their harassers more determined:

> **You4**: *I'm not sure. I think I would prefer people to not know that I'm using it. It could be ammo for people like, "Oh, this person, they're in Squadbox so let's try to overwhelm them," and then everyone starts sending emails. It's like they target people who don't want to be targeted, right? 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.*

> **Fan1**: *I think that would be a bad idea because if it's a person who's deliberately out to harass you, you want to give them as little feedback as possible on whether it got through or not. It's similar to how some forums do a thing called shadow-banning, where they ban someone, but that person can still make posts, and they'll see their post goes up when they're logged on, but no one else will actually see their posts. I feel like when it comes to trolls and harassment, there should be as little feedback to them as possible.*

Some interviewees mentioned that they thought the transparency of an auto-reply such as that would be good for sender privacy:

> **Act1**: *I think that that's really cool. Especially if it is like a Squadbox email address, like that second track you were describing where it's a public email that you put up that is moderated. I think that's really cool and transparent.*

> **You4**: *Privacy-wise, it's probably a better option to let people know that their email's going to be read by someone else.*

### Moderatoring for Others

We inquired whether interviewees would be willing to use Squadbox in the opposite way – as a moderator for someone else.

Most of them said they would be willing, at least for close friends. They viewed it as a tangible way to help and support others:

> **Jour1**: *I'd be happy to do that. 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.*

> **You2**: *Yeah, that'd be great. I'm used to it at this point, and I am very used to it coming at me and being about me. I think I have a pretty decent shield up about it at this point for myself, and even more so for other people. Especially because I understand what they're going through with it, so being able to protect someone else with an armor that I've build up for myself makes a lot of sense to me.*

> **Ex1**: *I know that having my friends make those offers [to mediate communication with my ex] for me were really valuable. I'm more than happy to return the favor if somebody needs it.*

Several interviewees were interested in a reciprocal relationship – moderating their moderator's emails:

> **Res1**: *Yeah, I would be willing to. Especially if it was like a mutual, reciprocal thing.*
>
> **You2**: *A moderator swap would be interesting.*
>
> **Fan1**: *This would actually be, this would be good for people who are ... A group of friends who are all equally likely to be harassed. It's emotionally painful to see the shit that your friend gets. That would suck. But if everyone in that small group of friends is a likely target, then everyone can review each other's.*

Others said that they would, but would need limits in place on when and how much:

> **Act1**: *I think that I would. I definitely would not be able to do it all of the time. I think that it would be a really wonderful thing I would love to do when I do have the strength to do it. I would feel very good just doing that, it'd be satisfying.*
>
> **Res2**: *Absolutely. I think it's because I feel pretty clear and safe myself I guess. I mean I would have a limit probably to how much I could do it. I'd be something you'd want to rotate, but I would also feel like I was doing at as a support for them, which would feel good. When somebody is in a situation like that, you want to do something to support them.*

Again, these interviewees mentioned how they would feel very good about helping people in this way.

For another interviewee, it depended largely on the individual and the type of harassment they received. They only felt that they could moderate for people whose harassment was very different than their own and wouldn't feel like an attack on them, too:

> **You3**: *I think so. Yeah. Depending on who they are and what harassment they get. If it was someone whose harassment was far enough removed from myself that I would feel comfortable reading it, then yes. I think reading anyone's hate comments can be draining because it's way too much negativity that no one should have in their lives. 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.*

You2 agreed that reading harassment feels different when it's not about you or aspects of your identity:

> **You2**: *Whenever I moderate my comments I'm like, I'm going to go into my comments now, I'm going to see if there's anything I need to delete, and then I'll go out of it and I'll be fine. I think that's a head space I could more easily get into if I were looking at someone else's comments, because it's not about me. It's about someone else's identity. It's hurtful to see those words directed at anyone else, but it's easier to stomach because it's not about me and my identity.*

Ex1 speculated about the same thing:

> **Ex1**: *It seems like that would be easier to do that kind of sorting if it wasn't for you. I can imagine that it's easier to just delete things that don't apply to you and not really internalize it the way you do when it's addressed to you.*

Ex2 said they would moderate for very close friends, but was rather hesitant to take on that kind of responsibility. Res1 also would for very close friends, but worried that moderating for someone they didn't know as well would be more difficult and error-prone. You4 (who was against having their own friends moderate for them, preferring the paid option) was decidedly against moderating for their friends:

> **You4**: *No, no. I love them, but no. No, that's not something I would ever willingly ... I can't think of a worst job to have.*

We inquired what kind of limitations those willing to moderate would want to be able to set on their moderation. Besides You3's limitation on the type of harassment they would moderate, all of the other limitations interviewees mentioned were time-based:

> **Act1**: *I would definitely only want to do it for a certain window of time. Just because it can get really exhausting. If somebody is getting an influx of emails, I feel like any more than a few hours can be really draining.*

> **Res1**: *Yeah, probably a time. I feel like this is the kind of thing that because you can work on anywhere on a computer, you could probably very easily get sucked into it, and just like do it all the time.*

### Desired Filters

We discussed with interviewees what kind of filters they would want to use to determine which emails need review, which emails should be automatically rejected, and

which emails should be automatically approved. The filters they proposed can all be categorized as either: 1) based on the actual words in and content of the message, 2) based on the sender's identity, or 3) based on the message's conversation context.

**Filters Based on Message Content** Multiple interviewees said that the ability to create filters based on words to determine what gets held for review or rejected would be helpful to them:

> **Jour2**: *If any of them have the "c-word" in it, that would be a good one. And then, you know, you could put in all of the anti-[religion] taunts. And then there are certain other words like "shrill," or ... If they're swearing, that would be a bad sign.*

> **You3**: *I would put in blacklist words. Like what I have on my YouTube channel and Twitter and stuff for slurs and stuff.*

Act1 suggested that such filters should not necessarily lead to a message being automatically rejected, but should just mean a moderator reviews it:

> **Act1**: *There are some words that I consistently know people will use to harass me and it's very easy to make a list. The only problem is sometimes people will use those words thoughtfully. I've had friends use those words, too, and with all of the moves to reclaim slurs... Especially things like slut and whore and cunt. You can use those perfectly fine, and if I set a Gmail filter and put all of those in spam, I could lose some emails that are worth reading, but having a filter system like with Squadbox where certain emails are flagged to be moderated, then I'm not losing those emails that I should be reading. But I can guarantee if somebody calls me, for example, a white bitch in an email, that's probably not an email I need.*

You3 reinforced that idea, stating that their word-based filter on YouTube occasionally had false positives:

> **You3**: *I also 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 as a topic of conversation. If I have the energy to, I'll go through my 'held for review' folder to look through those to see the ones that are actually not harassment that were filtered out.*

You3 also noted that word-based filters could often fail to catch words:

> **You3**: *The difficulty with the filters is that I may think of a word that I want to filter out, but then I forget a variation of the word. If there's a slur that I want to filter out I have to make it plural, and then also have to think about the common misspellings of it because people suck at spelling online, and I'll get all those words that did not go through the filter because they weren't spelled exactly how I had them in my filters.*

Similarly skeptical of word-based filters, You4 said,

> **You4**: *I can never create a list of words long enough to capture all of the things. I'll think I have all of the words down, and I post a new video and the first four comments are really creative ways of saying that I should kill myself. It's just like they manage to get around all the words. It does filter out a large chunk of them... but they're creative. Even then, sometimes with people in nice messages quote things, like, "Oh, someone said this to me," and they're not saying it to me, but it's in their message, and that gets filtered out, and then I feel really bad because I'm not responding to these nice messages because of it.*

So, while users wanted word-based filters, many were hesitant to have messages entirely rejected just based on the words in them. The difficulty of constructing and keeping up a list of words to filter suggests that it might be helpful for moderators to assist.

On the other hand, You3 said there were some messages they would want rejected without any human review. They could be certain a message was harassment if it contained a particular word/phrase, and they didn't want moderators to see messages with that word/phrase due to privacy concerns:

> **You3**: *I have [redacted] in my filters... If someone's helping me moderate, I would not want them to see that... I know that's not a message that would possibly be something I would want to see, and I wouldn't want anyone else to know what that is.*

Thus, there was interest in both having words lead to automatic rejection, and words lead to moderator review.

As an additional content-based filter, one interviewee also suggested coming up with some way to determine if the message subject and body's words matched up:

> **You2**: *I would want to see if there's a way to separate if the subject heading doesn't match the content, if they don't have anything to do with each other, because that would probably be a troll.*

And another interviewee suggested filtering based on whether a message has suspicious links:

> **You2**: *I think it would also be great to have a filter for anything that has links in it that's not to a known social media, because I've had people sending me links that I was supposed to click on them and then they would track my IP address and scary stuff like that, that would be great to filter out.*

Interviewees also wanted content-based filters for determining messages that should be automatically approved rather than moderated for privacy reasons:

> **You3**: *If I have someone helping me moderate who I don't know in person I would maybe have my address blacklisted in that setting, stuff like that.*

> **You2**: *I wouldn't want moderators to see anything that is financial.*

**Filters Based on Message Sender** Interviewees suggested several filtering mechanisms based on who the sender of the email was. One of these was a whitelist of trusted senders whose emails should be automatically approved:

> **You2**: *Anyone who I've contacted or who has contacted me previously, so contact list would be great. I guess it would be good to import a contact list from my personal email too, just in case anyone I know is trying to connect with me via that way.*

> **You4**: *I feel like all of the emails that I wouldn't want Squadbox to see are people I've already emailed or are in my contacts list or whatever, if they're from friends or family or people. Or I could just have a whitelist, and if I ask one of my friends to email me, I could just add them to the whitelist. These are people who are allowed to email me.*

> **Jour2**: *It would be good to add senders to a list that would get right through.*

A couple people thought that in addition to their own contacts, their friends' contacts could generally be trusted and not need moderation:

> **You2**: *I wonder if there'd be a way to not filter out anyone who's a referral, so if a friend or someone is mentioned in an email, so my friend told me to contact you, that kind of a thing would be great to just immediately get.*

> **Res1**: *It's just like how Facebook has that feature, like you can't send a message to someone unless they're your mutual friend? Unless someone in my contacts is shared with someone in their contacts, like that kind of thing. That would be one way. If there's no one in their contacts that is in my contacts, then it would automatically have to be reviewed.*

Interviewees also thought that data about the sender such as their email domain, location, and IP address could be useful ways to filter messages, or at least useful information to show them or their moderators for context:

> **Res1**: *If you can determine if this is someone who is part of my institution, or is it someone who's from another country? That kind of stuff I think. I don't actually know how helpful that would be, but it would be nice just to have that kind of stuff labeled already.*

> **Res2**: *For example anything with [redacted].edu would be... that could be really straightforward. Because everyone's got that here at my university and probably at least 70 percent of my daily correspondence is across campus.*

> **You2**: *Yeah, I think anything that's a website domain [as opposed to from Gmail, etc.], I would probably want to just go straight through to me, because that's typically more business contacts.*

**Filters Based on Conversation Context** Some interviewees felt that conversation context could also be a useful way to filter. You2 thought new messages that were part of an existing thread should bypass moderation. Similarly, Jour2 said:

> **Jour2**: *I'm not writing back to the ugly ones in general. So, maybe [threads I've replied to] should all go through.*

Res1 thought that in general messages after the first in a thread would not need moderation, but that the option for later ones to be moderated would be valuable:

> **Res1**: *Maybe like an option that if someone emailed me with that kind of message and I replied. And then it escalated, I could just kind of forward it to a moderator. I could say could you please take this off my hands sort of thing... You just have to like tag it for continual review.*

### Whitelisting and Blacklisting Permissions

We asked whether interviewees would want their moderators to be able to add to their sender whitelists, and their sender and word blacklists. Several did want help:

> **You1**: *Yeah, absolutely. The lists are always evolving so if they can be improved upon, that's great. Even if there's a word on there that is catching more neutral messages on accident, that sort of thing would be useful to know too.*

> **You2**: *Yeah, I would definitely want help. Even with my very long block list now, I'm still constantly adding stuff. People who comment and do this abusive crap, they know there'll be a block list, and they know when stuff starts getting hidden that they have to start spelling things funny or doing all this stuff to get outside of the filters. Yeah, it needs to constantly be morphing and I can do that, but also people who are reading my harassing emails would be able to do that as well.*

> **You3**: *Yes. I'm constantly finding more words that I realize I should have had blacklisted that I haven't had blacklisted and I'm constantly revising it. If I was not the person looking through the messages being filtered, I would want someone who is to be able to notice things that were missed by the filters to be able to edit it.*

Some interviewees wanted the moderators to be able to suggest additions or removal, but to have the final say before changes took effect:

> **Act1**: *I think it would be something I'd want to do in collaboration. I'd love to just take a look at it before it's implemented. But I do think that that would be a smart tool to have. Especially if they're seeing a pattern while moderating that I haven't seen because I'm not reading those emails as much, I think that that could be really useful.*

> **Jour1**: *I'd want them to suggest it, yeah. For me, I'd always want final veto power over that stuff. I wouldn't want anyone making those decisions for me.*

Ex2 agreed that they would want the final say over the lists, but would want moderators to be able to make suggestions. One interviewee, Fan1, was unsure, and suggested it should be an option for users to choose.

### What Happens to Rejected Messages

We asked interviewees what they would want to happen that messages that their moderator rejected – suggesting that for example these messages could be tagged, or directed to a particular folder in their inbox, or kept on the Squadbox website for future access. Alternatively, the messages could be deleted altogether.

One interviewee who felt reading their harassment (largely from people who disagreed with their columns) was important, and so they wanted them to be mixed in with all of their emails, but given a particular tag:

> **Jour2**: *I think, probably, sending it to me with a tag would be better, because ... I mean, I can't imagine ever saying, "Oh, let me sit down and check all of my hate mail. That would be a really good way to spend the next 20 minutes." But if I know what's coming, then I can decide, like, on an hour by hour basis, whether to click on it or not.*

Several other interviewees indicated that they wanted the messages to go to a separate folder in their inbox, and/or be available on an online archive. They liked this because then they could still read them when they wanted to, but not be disrupted by receiving them:

> **Jour1**: *I would want them to be put in a separate basket with a subject line that somehow indicates the subject matter and the severity of the content. Then I can look at them as I have time.*

> **You2**: *Probably, as much as that sucks, sometimes being able to see the harassment is helpful for my brain. I don't really know why, but it would maybe be helpful to have them just sort it off to the side somewhere. I think the categorizing and labeling would be key in that.*

> **You4**: *I wouldn't want to be notified every time that it happened. I wouldn't want a notification like, "Hey, you just got a shitty message!" But if there's a folder that I could check on every once in a while. Like checking your spam folder every few days.*

Act1 said that reading their harassment and having evidence of it was important to them, but they wanted it to be separated from their other emails, so they could read the messages by choice:

> **Act1**: *It can be very useful to know what I'm being sent if I'm in the right state of mind to read those messages. I wouldn't necessarily want them to pop up in my inbox because that still disrupts my day, but if there were a folder or filter where they would just tuck away. Or an interface somewhere where I could be like, "Okay, it's six o'clock. I'm about to go home. Time to check my terrible folder first." And could schedule it. It feels like I'm also choosing to interact with that, too. It gives me control back and I do want to see what's in those messages. I'm a control freak about my harassment. I want to know what people are saying, I want to know how the conversation is changing. I don't want those to disappear into the ether, either. I also find them super useful when I'm writing about harassment that I can pull up an example and be like, "This is an actual thing someone said. Here's a screenshot." That can be tremendously helpful to have.*

You1 similarly wanted the messages stored away as a way of documenting their harassment:

> **You1**: *I would want them to go to a separate folder that I don't need to look at until I need to look at it. If I ever wanted to help someone like yourself who's doing research about this stuff. Or YouTube sometimes contacts me about this stuff. Just to have it there. I wouldn't want to get rid of it. But I don't want to see it either.*

Res2 stated that they did not want the rejected emails sent to them, only the explanations for rejection, and they would want the emails to be accessible to them on our website.

Ex1, on the other hand, felt the messages from their sole harasser could be deleted altogether, and wanted no notification about them or archiving: 'There's nothing they need to say that I need to hear at this point.'

### Information from Moderators About Their Decisions

We asked what information interviewees would want to get from their moderators regarding their decisions – whether, for example, they might want a summary of what a harassing message was about, or have messages tagged with certain characteristics.

Most interviewees expressed strong interest in having tags, and some were interested in summaries as well. They felt these would help better prepare them to read harassing messages, or help them to decide whether they wanted to read them at all, by giving them more information about the content. They also felt tags could be

used to give them characteristics like the priority of a message, or whether it needed a reply:

> **Res1**: *Probably just like two or three sentences. Like what kind of email this is, and like one specific reason as to why it might be filtered.*

> **Act1**: *I think that a summary would sometimes be good. Because very frequently, I do find it useful to know what is in a harassment email that I'm getting. For one thing, it's interesting as an Act1 to know what people are attacking me for, but if it's somebody who's just being a creepy stalker, it's nice to know this email contains a reference to my dad, or this email contains a reference to this email address. It's nice to have somebody do the Cliffs Notes version of the horrifying email. But even if just a thing that says rejected; contains slurs. Just like a brief reason why it was rejected can be really helpful if it's not otherwise important.*

> **You1**: *I think tagging things would be helpful particularly when I'm making videos regularly and there's lots of email coming in. And some of the emails are more urgent. Or the writer is very stressed out. Tagging them to bump them to the top. And then tagging potential business stuff as well. Separating that out, and tagging things that don't necessarily need a response.*

> **You2**: *Yeah, email has those little label things, that'd be great. Just saying oh, this is a fan that you should talk to, or this has some word and just got filtered through accidentally.*

> **You3**: *Yeah, if an email was not just entirely trolling, slurs and stuff, I think I would want a more toned down summary of what it was about to decide if I actually wanted to read it or not. I think that would be useful. Tags would also be useful as a way of trigger warning, I guess since that's not something people usually put on emails.*

> **You4**: *Just being tagged with the general reason they were red flagged, like death threat or just mean or something like that. I think the vaguer it is, the better. The more specific it gets, the more I don't want to know about it.*

> **Jour2**: *It would be super helpful to have some preview, beyond a tag. You know, even if it's just a few words like "sexist," or "racist," or, you know, "swearing," or "personal attack," or "threat." That would be helpful.*

Fan1 suggested that there should be a preselected list of tags, to make this work easier for moderators:

> **Fan1**: *Tags would actually be really helpful. Spam, or abusive. I'm thinking probably to make it easier for the moderators, just have a short list, multiple choice for why this thing was rejected. It's spam, or it's harmful or, whatever other options you could think of for why an email would be rejected.*

And Jour1 suggested a rating system for how severe a message is:

> **Jour1**: *I'd want it to be categorized somehow and rated as far as the severity of it. Like, this falls into the general category but on a scale of one to five, with five being the most harassing, it's a four. Or this is about the story that you're covering, it's borderline, it's something immediately that you should look at, but as far as the severity of it it's more like a two.*

Res2 said that they would want to see the moderator's explanation for their decision, because it would provide them with an opportunity to correct the moderator's rationale in the case of a miscategorized email.

The interviewee who was targeted by their ex-significant other noted they wouldn't need any information from moderators, because they wanted all emails from that person to be rejected unless they were about events both of them might attend:

> **Ex1**: *Because the parameters are so strict of what I need to see, I don't need them to tell me that it's about an event, because that's really all I should be seeing.*

### Emails That Need Escalation to Law Enforcement

We asked interviewees what they would want to happen if a moderator reviewed a message and found that it contained something (such as a death threat) that possibly warranted contacting law enforcement. Would they want to be notified? Would they want their moderator to contact law enforcement, or would they want to do it themselves?

Regarding being notified, interviewees suggested categorizing these messages separately and receiving immediate notifications about them from the moderator:

> **Jour2**: *I think they need a special tag, that we should probably act on them urgently.*

> **Fan1**: *I would definitely want my friend to tell me that so I can take precautions.*

**Act1**: *I think what I would probably like is an alert sent to me, either they would text me or maybe there's some sort of automated thing where I find out right away but I don't necessarily have to read it myself yet. I can elect to read the message, the email, if I want to, but it's not beings sent to me just itself, it's like a friend being like, "Hey, we just got this threat or this dox of your personal information. Here's the message if you want to click into it, but we should escalate this." That would be super helpful.*

**You1**: *Yeah, I think those should go into their own category to file a report. I would like to have those be sorted out separately from the rest of the harassing email.*

**You3**: *That would be something that I would want to know about. I wouldn't necessarily want a mod to deal with on their own. I've gotten messages about people threatening to dox me, and that would be something that I would want to communicate to the people that I'm living with.*

Res2 also said that they would like a notification with some details about the message in terms of it being a credible threat – whether the sender had their address or knew their workplace, for example, or was just sending something angry.

Regarding whether the moderator should report the message to law enforcement, or if they wanted to do it themselves, interviewees had varying opinions. Some of them felt it was an action they should take themselves:

**Jour2**: *I'm not sure whether the moderator should do the acting or whether the moderator should kick it back to me to take action on it. I guess I prefer the latter. I mean, it's hard because I don't want to see them, but if I don't see them I don't know what I'm up against.*

**Res1**: *I would want to personally inform the police myself. I wouldn't want someone doing that for me.*

**Ex1**: *I would want to do the notifying... If they're saying things that justify law enforcement, then I'm quite happy to be the one to run those in.*

While others wanted to be spared from reading the threats and contacting police:

**You2**: *I definitely think death threats would be great to automatically go to police. Me having to go the extra step of reading the email through and being traumatized by it, and then me having to call the police separately and having to go through that interaction... It's just a lot of work.*

> **You4**: *If they could deal with that, that'd be awesome. It would just save me from having to do that.*

A few interviewees expressed skepticism about whether someone else would actually be able to file a police report on their behalf:

> **Act1**: *I imagine it would probably be hard for them to take action without my involvement because of the way the police are.*

> **You1**: *If it's something that someone else can do, that would be awesome because it's upsetting to deal with. But if not, then to me.*

### Turnaround Time

We asked what kind of turnaround users would want on moderated emails - i.e. how long it takes emails to be handled by their moderators. Their answers ranged from 30 minutes to a couple of days, depending on what type of messages they wanted to go through moderation, the type of harassment they received, and whether the email account was personal or for work. The most common answer was 24 hours.

> **You4**: *I would think 24 hours at the latest. 12 would be good. I don't know what's realistic. Probably 12 hours would make sense because if you were waiting longer than a day, you could miss a deadline for something.*

> **You3**: *Since I don't get a large number of emails, I think a day. There's never usually something that I need to respond to quicker than that.*

> **Act1**: *I think if I were going to use this on my personal email address, I wouldn't need super quick turnaround. 12 hours to a day would probably be fine. Especially if this is something being done by my friends.*

Some users said even longer than that was fine; as Ex1 only wanted emails from one specific sender moderated, and was not concerned about threats or urgent messages from that sender, they said 'a couple days would be fine.' Res1 also said that a couple of days would be good.

Some interviewees noted that they would want a shorter turnaround time if using Squadbox with an account where they received professional emails:

> **Act1**: *I think that it depends on the account. If somebody were using this for a work email address, that's where it gets sticky because you really do need emails quickly.*

> **Jour1**: *Pretty damn quick. It depends on the topic. I would like some kind of way to rank them by importance. If it's about something I'm covering right now, or a story that I recently wrote, I'd want to be flagged quickly, I'd want that to be reviewed quickly. If it's something, just a general thing like, "I hate your work and you suck," then that's something that I think can take more time. If it's possibly about something that I need to decide whether it deserves a correction, for example, those are things that legally, as a journalist, you can open yourself up to more liability if you don't quickly address it.*

Though one interviewee, perhaps due to their working independently or their own personal email habits, did not feel the need for quick moderation of work-related emails:

> **Fan1**: *If it's professional I could go up for the day. If it's just professional-ish emails. That's basically my email reply schedule anyway, is about 24 hours at the minimum.*

Jour1 mentioned that their desired turnaround would largely depend on the type of the email – whether it was something urgent and work-related, or just a harassing message. Spoof1 shared that sentiment, noting that 24 hours would usually be fine, but they would want something more like 12 for important emails. Ex2 said that they want 30 to 60 minute turnaround if they were currently receiving threats, as those might have contents needing to be urgently dealt with.

# Chapter 4

# System Design and Implementation

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. In this chapter, we summarize the needs and goals we identified, followed by user scenarios inspired by our interview subjects, followed by features and technical implementation details.

## 4.1 User Needs and Design Goals

Taking together all of the previously described findings, we identified several user needs that current platforms do not address. We label each need with a short word/phrase so that we may reference them throughout the chapter:

- **DIVERT**: Users need to be able to divert harassing messages from their inbox or platform equivalent.
- **DoS** (denial of service): Users need to be able to maintain private and public communication in the face of harassment

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

- **ON/OFF**: Users may need to ramp up or down mitigation strategies as harassment comes in waves

- **INSPECT**: Users at times need to be able to read or get an overview of their harassing messages

- **COLLAB** (collaborative filter): Users need help managing blocklists and filters over time

- **DOCUMENT**: Users need help collecting and documenting harassment for official reports

We additionally determined several design goals necessary for successful tools for friendsourced moderation:

- **CUSTOM**: Tools need to be customizable to suit a variety of user needs and preferences. Subjects described different preferences for what actions they wanted moderators to take and what powers moderators should have.

- **PRIVACY**: Tools should allow users to mitigate privacy concerns. Many subjects had messages they preferred to keep private, even from friends.

- **MODWORK**: Tools should effectively coordinate moderators and minimize their workload. While subjects and their friends were eager to moderate, recipients expressed guilt about asking for help and the potentially high volume of messages.

- **MODTRAUMA**: Tools should minimize secondary trauma to moderators. Subjects expressed concerned about the emotional labor their moderators would do.

We now turn to describing how Squadbox works and how its features align with the identified needs and goals.

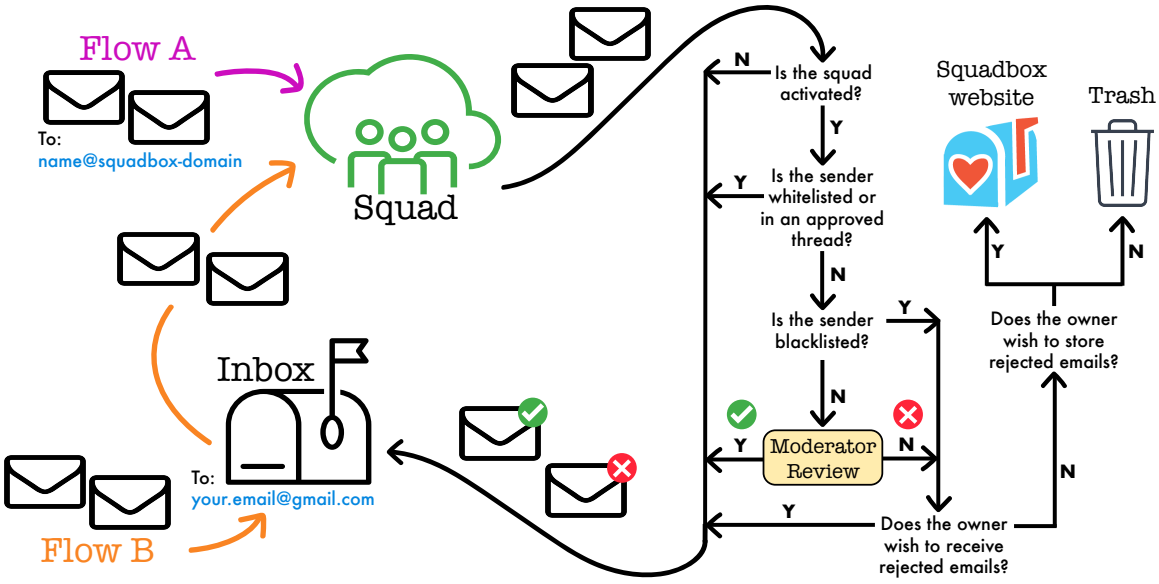


Figure 4-1: Diagram of the flow of messages through Squadbox, including Flow A, which allows users to have a public moderated account, and Flow B, which allows users to get their current account moderated. From there, various settings define whether emails get moderated and where they go.

## 4.2 User Scenarios

Here we describe potential user scenarios, with the corresponding flows of emails shown in 4-1.

### 4.2.1 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 (DoS).

### 4.2.2 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 (DoS), 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, along with adding a forwarding filter on her email client specifying that only strangers' emails go to her moderators. She can also turn on Squadbox when she starts getting harassment but then turn it off when it dies down (ON/OFF). 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 (INSPECT). Eve's use of Squadbox makes it possible for her to use her current email account more effectively by filtering out the unwanted content while still giving her access to the important messages.

## 4.3 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.

### 4.3.1 Features for Reducing Moderator Load and Increasing Privacy

To begin, we describe automated moderation features that work to reduce the burden placed on moderators (MODWORK) as well as support increased owner privacy (PRIVACY).

### 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 (COLLAB, CUSTOM). Finally, we develop tools to easily import from one's contacts and export to filters. We give owners the option to "blacklist" senders whose emails should be automatically rejected. A planned improvement to this feature is allowing owners to specify that rather than being automatically rejected, blacklisted senders' messages should be held for moderation, and all other messages are automatically approved. As with whitelisting, this feature aims to reduce the workload for moderators (MODWORK). As with whitelists, owners can choose whether their moderators can add to their blacklists. Such filters partially alleviate any concerns about slow moderation turnaround time, and helps owners feel more in control over what messages their moderators see (PRIVACY). 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 let through 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, or manually disable follow-up moderation on the sender-thread level if they have this setting turned off. This feature provides more fine-grained control over how much of conversations moderators can see (PRIVACY), reduces the number of messages moderators must review (MODWORK), 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 (ON/OFF). To better accommodate this, users can deactivate a squad so that all emails will be automatically approved, reducing moderator workload (MODWORK). When it is reactivated, all previously defined settings, whitelist, etc. take effect again.

### 4.3.2 Features for Reducing Secondary Trauma to Moderators

Now, we describe existing and planned Squadbox features that work to minimize secondary trauma to moderators (MODTRAUMA).

#### Control over Viewing Harassment

Subjects described how receiving harassment in their inbox disrupted their day-to-day (DIVERT); 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 [17] 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. Finally, we plan to provide

training and support resources to ensure moderators are equipped to do their work and can get help if needed.

### 4.3.3 Features for Giving Moderators Context and Information

Next, we describe features that give moderators more information to better tailor their decisions (CUSTOM) and make moderation easier (MODWORK). These are shown in 4-2.

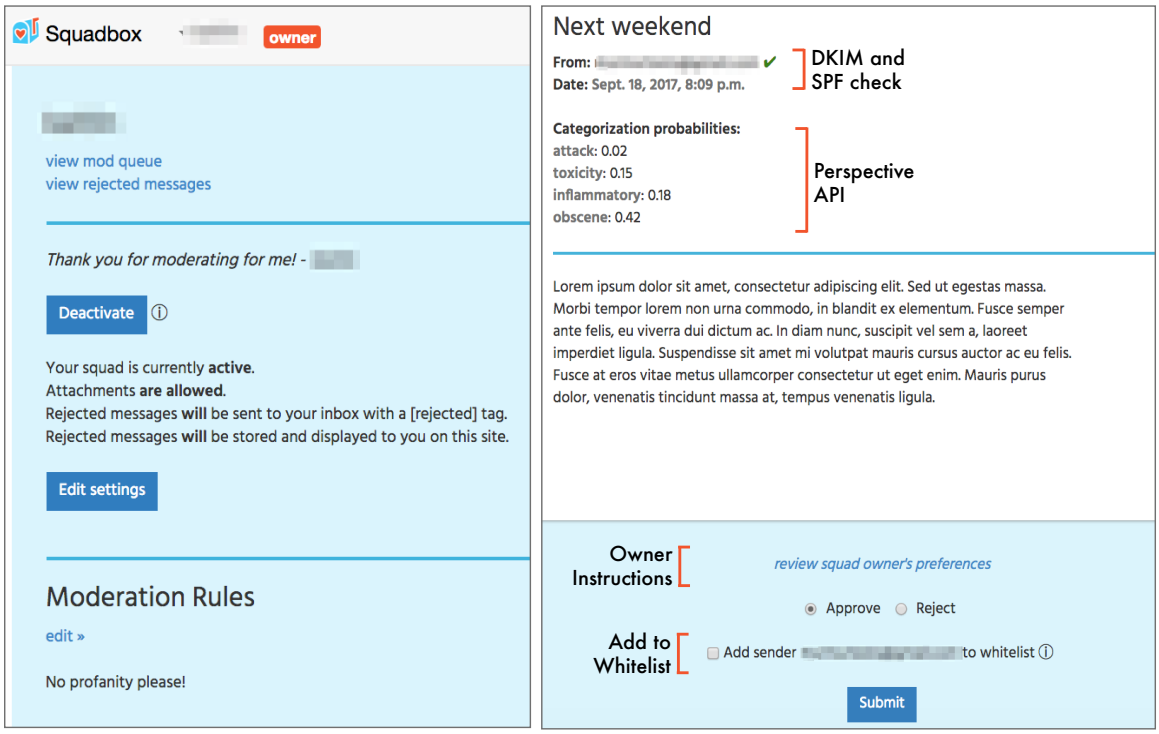


Figure 4-2: On the left, an owner's view of the information page for their squad. On the right, a moderation page for the moderator.

#### 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 [29] or have different actions they want moderators to take, we allow owners to give instructions to their moderators via a freeform text box (CUSTOM).

### 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. While we do not make any automated decisions based on this as it is possible for non-malicious emails to fail these checks if they are incorrectly signed, we show the results of the verification to moderators in the interface to help inform their decision. 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. In the future, we plan to add additional features surrounding this verification process, such as allowing the owner to require moderation on whitelisted senders if their verification fails.

### 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 [17]. These scores are shown to moderators when they review messages. In the future, we plan to expand the integration by automatically generating tags based on scores, and possibly try prioritizing messages in the queue based on their scores. While none of these features automate the moderation process, they aim to assist the moderator and ease their workload.

### 4.3.4 Features for Giving Owners Customization Capabilities

Finally, we describe features that allow owners to customize what should happen to harassing messages (CUSTOM).

#### Divert and Collect Harassing Content

We give owners the option to receive harassing content (INSPECT) or file them into a separate folder (DIVERT), 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. If the user has chosen 1), we will tag the message and deliver it to them. If the user has chosen 2), rejected messages will be stored in the Squadbox database. The user can then go to a page on the website and browse an index of the rejected messages, consisting of message metadata and moderation information such as who rejected the message and their explanation for rejecting it. They can then choose whether or not to click on and view the body of the message. We provide downloadable Gmail filters for owners to automatically forward emails with a "rejected" tag into a separate folder.

### 4.3.5 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 read the subject lines to decide whether to open messages, or alternatively add a filter in their mail client to customize where messages with those tags go. Messages can also be grouped or sorted by tag on the website (DOCUMENT).

### 4.3.6 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 (INSPECT). 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.

## 4.4 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 here two different ways the system can be used, as well as optimizations and extra features available for Gmail users.

### 4.4.1 Flow A

**Everything goes through Squadbox. Some of it gets automatically approved, and the rest is moderated.** This flow works like a moderated mailing list with one member and one or more moderators. When a message arrives, we determine if it can be automatically approved or requires moderation. If it is automatically approved, it is immediately sent on to the intended recipient's inbox. Otherwise, it is stored on the server until a moderator reviews it, at which point the message is either delivered or not, according to the moderator's decision and the owner's preferences.

This flow is simpler design-wise and will work with any email client that supports simple forwarding. However, it requires the tradeoff that all mail passes through the Squadbox server, even if it isn't read or stored.

### 4.4.2 Flow B

**Only mail with specific characteristics goes through Squadbox, and those messages are all moderated.** 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: in English, that filter's rule is "forward messages that don't have `[secret hash]` in the `list-id` header field". We need this capability to prevent a forwarding loop—by slightly modifying messages that pass through Squadbox to no longer match the filter, we stop messages Squadbox has already seen from being re-forwarded. The `address X` used in the forwarding filter contains a secret hash, known only to the owner and Squadbox, to make it harder for attackers to falsify an approved message by setting the `list-id` header themselves. If the address gets compromised, for example if the owner forwards an approved email to an unsafe sender or their email account is compromised, the user can generate a new address and filter.

In this flow, messages from whitelisted senders or that are otherwise automatically approved are immediately sent back (with the list-id modified) when Squadbox receives them; the rest are stored on the server until they're moderated.

The capability to selectively forward based on a filter is common in email clients (Gmail, Thunderbird, Apple Mail), but not universal. In the future, we will provide modification options beyond changing the `list-id` in case users want to preserve that field's original value or if their client does not support filtering on it. We will also provide tutorials and methods to more easily set up such filters.

Overall, the selective forwarding approach better addresses users' privacy concerns about exposing their inboxes to the system or their friends, but it requires more complex setup and forwarding capability.

### Flow B - Optimizations for Gmail

For Gmail users, we leverage the Gmail API to add several optimizations that reduce the Squadbox server load, mitigate privacy and security concerns, and enhance the user experience:

**Whitelist Creation** As shown in 4-3, Gmail integration makes the whitelist creation process significantly easier for the owner by importing their contacts and people that have emailed them in the last three months. We display a list of names and addresses to the owner, grouped by Gmail "smart categories", and allow them to deselect addresses individually or by category. We also provide a slider that automatically sets a cutoff based on their frequency of contact with the user.

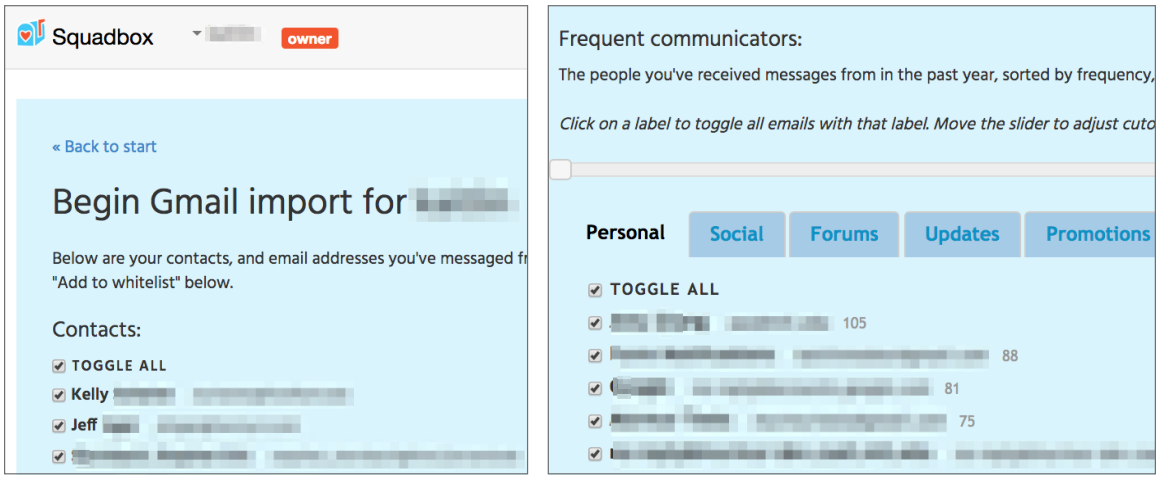


Figure 4-3: Squadbox generates whitelist suggestions from an owner's Gmail contacts

**Filter Creation** The Gmail API allows us to automate the forwarding setup for the owner by programmatically creating filters. As demonstrated in 4-4, we use Gmail's rich filtering language to express the following: "if an email is not from someone on my whitelist and doesn't have `[secret hash]` in the `list-id` header field, forward it to Squadbox and move it to my trash." This has two advantages: first, it cuts down on the volume of email that actually must pass through Squadbox, since messages from whitelisted senders stay in the inbox rather than take a round-trip. Second, it gives the owner a greater sense of security knowing that Squadbox and moderators will never see messages from certain senders, as they are never forwarded. Finally, we automatically refresh the hash in the filter continuously to protect against attackers.

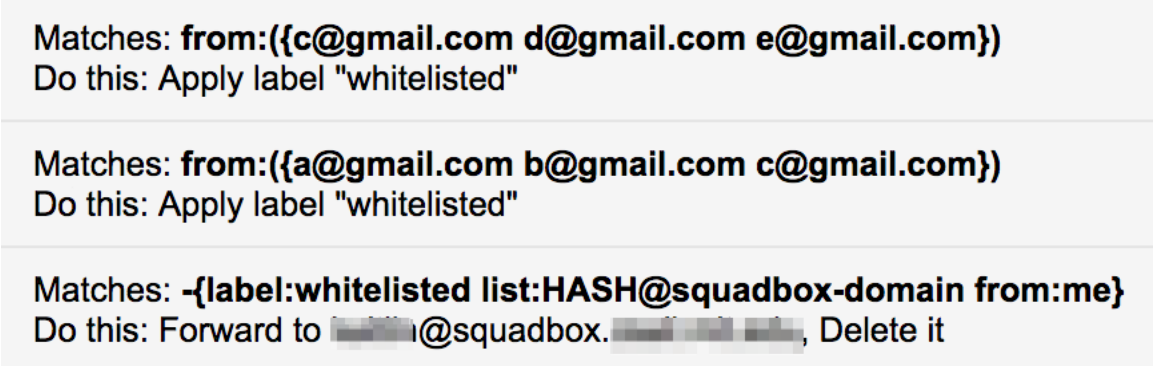


Figure 4-4: An example of Gmail filters for Squadbox

**Delivering Messages To The Owner** When an email is approved, we use the Gmail API to move the deleted email from the trash back to the inbox. This has the advantage of giving the user back the exact original message, rather than one which has passed through the server and been slightly modified. If the API call fails for whatever reason, or if the user has not granted us sufficient Gmail permissions, we default to emailing them a new copy of the message, as in the typical Flow B. Since rejected emails have additional data (a rejected tag, and reasons for rejection), for those we instead send a new email to the owner, with the "From" address set as Squadbox, a "[rejected]" tag in the subject line, and the provided reason for rejection at the top of the body, followed by the body of the original message. We also put the original sender address in both the start of the subject line and in the `reply-to` email header. We also provide the owner with a filter to move the messages with the rejected tag to a separate folder.

# Chapter 5

# Field Study and Demos

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 both owners and 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 motivate 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 ask their moderator to reject unwanted emails. For our test subjects, these unwanted emails were mostly spam and advertisements.

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

## 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

| 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 5.1: 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.

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.

## 5.3 Field Study Results

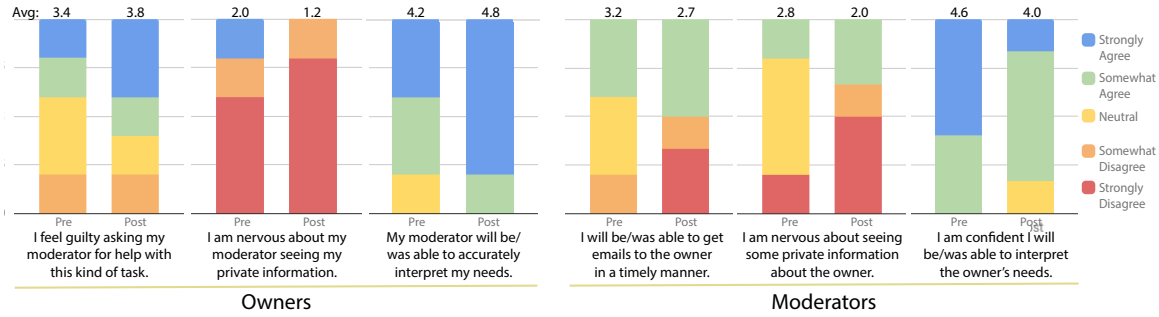


Figure 5-1: Comparison of agreement with statements before and after the field study, where 1=Strongly Disagree, 5=Strongly Agree.

**The whitelist/blacklist feature was an effective way to separate out potentially unwanted messages.** As shown in Table 5.1, 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 had used the example instructions shown in the Squadbox interface ("no profanity") rather than writing their own. 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, the rules they provided were all rather short (two 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 5-1, 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, with owners dropping from 3.2 to 3.0 and moderators from 3.2 to 2.7. 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. One owner suggested a way to alleviate moderator fatigue: implement a timeout feature, where if a message is not moderated for some set amount of hours (implying the moderator is busy), the message will be automatically approved. Finally, a common pain point among owners was that they occasionally received emails from new addresses that they wanted to see immediately, like password reset links or order confirmations. Although we mentioned to owners that they could see and also approve emails in the moderation queue on the Squadbox site, their frustration suggests that we should make clear that they always have access if they need.

**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 5-1. In addition, owners felt increasingly guilty as the study went on.

# Chapter 6

# Discussion

## 6.1 Field Study

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.

## 6.2 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 [20] 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.

As a counterpoint, three of the YouTube personalities we spoke with preferred paid, vetted stranger moderators, as their harassment was directly related to their jobs and source of income. For instance, one person said, "*Because this is my business, I don't want to ask my friends to basically do work for me, or a volunteer to do work for me. This is my full time job, so just as a business ethics thing, I would use the paid moderators*" [You1]. Despite their interest in paid moderators, You3 and You4 said this would not be financially possible yet, and You1 had been unsuccessful in hiring a moderator to date. However, these interviewees felt it would be important for the moderators to be vetted and for them to have an on-going business relationship where trust could be established, as opposed to anonymous crowd workers.

## 6.3 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 final option would be to continue using our email pipeline and use platform APIs to route content to email. This however, would require owners to access and perhaps conduct their platform conversations through email, which presents new usability challenges. 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.

# Chapter 7

# Conclusion

## 7.1 Limitations and Future Work

In our implementation of Squadbox, we encountered some issues with rate-limiting in the Gmail API, as well as issues in the non-Gmail case where clients believed Squadbox was spoofing headers because our server did not belong to the sender's email domain.

In the future, we hope to move to re-implementing Squadbox as an IMAP client, so it can fetch email from any IMAP server and easily move email between folders using the IMAP protocol instead of relying on individual email client APIs. Since multiple clients can access the same IMAP server, owners could still use whichever email client they prefer.

We also hope to connect Squadbox to the APIs for other communication platforms to enable moderating their messages, in order to make it more effective for people who are harassed across multiple platforms.

Finally, while our field study explored the use of Squadbox as a friend-moderation tool for email, it did not actually 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 takes an emotional toll. In future work, we aim to move forward cautiously with Squadbox, including giving more demos to

different harassment recipients as well as their potential moderators before initiating a small-scale release.

We are working with the Mozilla Foundation to help transition from Squadbox from being just a research project to a viable open source project. We hope this will assist us in finding people to contribute ideas and technical work so we can implement the several aforementioned features and improvements, and build a robust system users can rely on.

## 7.2 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 them from harassing messages. Building on this strategy, we propose friendsourced moderation as a promising technique for anti-harassment tools to better facilitate. 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.

# Appendix A

# Survey and Interview Questions

## A.1 Survey Questions

We indicated at the beginning that all questions were optional, and none of them were required to be filled out to submit the form.

1. Name
2. Age
3. Gender (multiple choice)
   - ☐ Female
   - ☐ Male
   - ☐ Other: _____
4. Which of the following categories describe you? (check all that apply)
   - ☐ White
   - ☐ Hispanic, Latino, or Spanish origin
   - ☐ Black or African-American
   - ☐ Asian
   - ☐ American Indian or Alaskan Native

- ☐ Middle Eastern or Northern African
- ☐ Native Hawaiian or other Pacific Islander
- ☐ Other: _____

5. What is your current occupation?
6. How many email accounts do you regularly (i.e. at least once a week) use? (multiple choice)
   - ☐ 1
   - ☐ 2
   - ☐ 3 or more
7. If you have more than one account, how do they differ? (for ex. "One for work and one for personal")
8. In what contexts do you use email? (check all that apply)
   - ☐ Work
   - ☐ School
   - ☐ Personal communication
   - ☐ Other: _____
9. In what way(s) do you typically access your email? (check all that apply)
   - ☐ Web browser
   - ☐ Desktop application
   - ☐ Mobile application
   - ☐ Command line
   - ☐ Other: _____
10. What email client(s) do you use? (check all that apply)

- ☐ Gmail
- ☐ Outlook
- ☐ Apple Mail application
- ☐ Mozilla Thunderbird
- ☐ Yahoo! Mail
- ☐ Other: _____

11. To your knowledge, is your email address publicly available on the internet? (multiple choice)

    - ☐ Yes
    - ☐ No
    - ☐ Unsure

12. If yes to the previous question, where? (check all that apply)

    - ☐ Social media profile
    - ☐ LinkedIn
    - ☐ Personal website
    - ☐ Website for your business/place of work
    - ☐ Other: _____

13. Does anyone besides you have access to, manage, or use one/all of your email accounts? (multiple choice)

    - ☐ Yes
    - ☐ No

14. If yes to the previous question, who? (check all that apply)

    - ☐ Assistant or secretary
    - ☐ Family member

☐ Other: _____

15. Estimate what percentage of the email you receive is unwanted (either harassing, or just irrelevant to you): (multiple choice)

    ☐ 10% or less

    ☐ 25%

    ☐ 50%

    ☐ Over 75%

16. How long have you been experiencing online harassment? (multiple choice)

    ☐ Less than one month

    ☐ 1-6 months

    ☐ 6-12 months

    ☐ Over a year

    ☐ Over 2 years

17. Have you been harassed for any of the following aspects of your identity? (check all that apply)

    ☐ Political opinions

    ☐ Race

    ☐ Gender/gender identity

    ☐ Religion

    ☐ Sexual orientation

    ☐ Appearance

    ☐ Occupation

    ☐ Disability

    ☐ Medical condition

☐ I have not been harassed for any of these

☐ Other: _____

18. Have you experienced any of the following types of harassment? (check all that apply)

☐ Sexual harassment

☐ Unwanted romantic interest

☐ Offensive or derogatoy language, slurs, or profanity

☐ Stalking

☐ Physical threats

☐ Being sent pornography/explicit content

☐ Threats to reveal your private information online

☐ Having private information revealed online ("doxing") or otherwise

☐ Related to a personal relationship (friend, significant other, ex)

☐ Emotional manipulation

☐ Extortion, blackmail, threats for money or other goods

☐ Disturbing images

☐ Manipulated (i.e. Photoshopped or otherwise) images of yourself

☐ Threats/insults to your self, ego, intelligence, or professional competence

☐ Other: _____

19. What platforms/methods have you experienced harassment through? (check all that apply)

☐ Email

☐ Facebook

☐ Instagram

☐ Twitter

☐ Reddit

☐ YouTube

☐ Online comments sections

☐ Text message

☐ Phone

☐ In-person

☐ Other: _____

20. Thinking back to the worst days/"peak" in terms of your online harassment, around how many messages did you receive? (multiple choice)

☐ Female

☐ Male

☐ Other: _____

21. On average, how many harassing messages do you receive? (multiple choice)

☐ 50+ a day

☐ 10+ a day

☐ 1+ a day

☐ Around 1/day

☐ Around 1/month

☐ Less than 1/month

## A.2 Interview Questions

Questions are broken down into sections by their general topic/area. The time in brackets indicates approximately how much time we tried to spend on the section.

### A.2.1 Email usage [5 min]

- What platforms do you use for online communication? Where does email fit in?
- Can you describe how you organize/manage your emails?
- Sense of email load: are you swamped with email, or is it rare/desirable?
- What things might you like to outsource to others in terms of dealing with your emails? For example, answering common questions or requests, creating to-dos
- How important is it for strangers to be able to reach you on your email?
- Can you estimate how much of your email communication is with frequent contacts vs. unknown/new contacts? How often do strangers contact you?

### A.2.2 Harassment [25 min]

- Think back to the last time you received a harassing message. [5 min]
- Can you describe the circumstances?
- What was the message about?
- What was that experience like? How did you feel?
- Do you know who the message came from? How do you think the sender got your email address/contact info?
- What account did you receive this harassment on? Was it a work email or personal?
- What did you do, if anything, in response to the message?

- Is there anything you wish you could have done about it? Any tool you wish you had?

### A.2.3 Defining harassment [5 min]

- What sort of things do you think constitute harassment?
- What level of language or expressed actions would upset you/would you like to banish from your inbox?
- Does your definition of/threshold for harassment differ for friends/people you know and strangers, or for personal vs. work email?
- Do you think you could define harassment in terms of a set of rules? Certain keywords, sending patterns, people, or any other feature?

### A.2.4 Overall experience with harassment [8 min]

- How frequently do you experience harassment? Can you describe how the frequency varies? (For example, are there short bursts of it every so often? More of a constant stream?)
- What, if, anything, do you think has triggered the harassment? (Sharing your opinions online, personal conflicts, etc.)
- Who is harassing you? Is it all strangers, or also people that you know?
- How motivated do you think your harassers are? Are they creating multiple accounts? Hiding their identities? Working their way around defenses you have established?
- How many different people (roughly) would you guess have harassed you? Do you think it is repeat offenders or usually different people?
- Is the harassment typically apparent right away? Or does it more often develop out of relationships or conversations that turn sour?

- What are the consequences of the harassment in your life? How disruptive is the harassment? Is it just an annoyance, or does it actively harm your life in any way? Is the volume the main issue? Or is it more about the emotional/mental consequences?

- Has harassment changed the way you communicate online, or the way you use the internet in general?

### A.2.5 Responding to / dealing with harassment [8 min]

- How do you generally respond to harassment, if at all? Actually replying to the message? Blocking the sender? Reporting the abuse?

- How do you prepare for / protect yourself against future harassment? Are there specific tools or processes that you use?

- Have you tried any actions such as blocking or blacklisting specific people, reporting the abuse, etc.? Did this work?

- Have you asked friends, family, or other services (law enforcement, hotlines, websites) for help? How did that work out?

- Have you taken any other actions in response, such as changing your email address, making a new social media profile, making an account private, or deleting an account?

- How do you think platforms could help you deal with harassment?

- Are there changes to the interface or additional tools that you think could help?

- If there was a group of people that were willing to help you deal with harassment, what do you think they could do? What areas would you like assistance with?

### A.2.6 Introduce Squadbox [5 min]

Squadbox is a tool to combat online harassment via email. It places a âĂIJsquadâĂİ of moderators in between you and the people who email you. Your squad is made up of one or more moderators, who review the messages sent to your Squadbox and decide whether the messages should be approved, rejected, or flagged, according to your preferences. The messages you want to see are then sent on to your inbox.

### A.2.7 Squadbox-specific questions [30 min]

- Ask if they have any questions/clarifications
- Ask for initial thoughts/reactions to the idea.
- Which of these two ways do you think you would use (both is ok)?
- What sort of filters would you want to set regarding what emails go to your Squadbox? Emails from any strangers? Emails with certain words you select?
- What sort of filters would you want to set regarding what emails *never* go to your Squadbox? (That is, emails you do NOT want your squad to review.) Emails from certain people you set? Certain terms?

**Friend-Moderators**

- How would you feel about having a friend/friends make up your squad?
- Do you think you have friends who would be willing to do this, and have enough time to?
- Do you have any concerns with regards to asking friends? Both in terms of your own privacy, and in terms of their well-being, your relationship with them, etc.
- How would you feel about asking this of your friends?

**Paid Strangers**

- Are you familiar with Mechanical Turk / crowdsourcing platforms? (If not, explain.)
- How would you feel about having a paid moderator(s) (from Mechanical Turk, etc.) be in your squad?
- Would you be willing to pay? If so, how much? (for ex,. 5 cents per email)

**Volunteer Strangers**

- How would you feel about having stranger(s) who volunteer be in your squad?
- Can you think of people or groups that would be interested in volunteering for such a service?

**Reverse - interviewee as moderator**

- Would you ever be willing to moderate for friends that ask you? What concerns would you have?
- Would you ever be willing to moderate for strangers that ask you? Either paid or volunteer. What concerns would you have?
- Would this kind of work be draining at all or upsetting?
- Would you want to set limits such as how many emails, how much time you dedicate?
- Would you ever want to step away from the task or pass it on?

**Privacy concerns**

- Do you have privacy concerns regarding friends reading the emails you receive? Specifically: emails from people you know? What about emails from strangers?
- Do you have privacy concerns re: reading the emails that your friends or strangers receive?

**Possible mitigation techniques**

- What if your emails were automatically made anonymous before moderator could see them – in no way attributable to you? An example of this:

  Hi NAME,

  Here's my cell number so we can be in touch the day you come to drop off the car - NUMBER.

  We have a tandem driveway so I'll need to coordinate with my two roommates so they can move their cars out of the way. NAME said you're going to drop it off on DATE right? Let me know closer to the date when you have a time frame of when you'll drop it off. My address is ADDRESS.

  Thanks! NAME

- Alternatively, what if your email were broken up into sentences, and each moderator could only see one sentence?

- Do you have any other ideas about how to improve/address privacy concerns?

- Which of the previous options would feel most private to you?

- How do you think the anonymization techniques would affect the ability of moderators to understand whether or not an email was harassment?

**Moderation technique/abilities**

- What if you had an automatic reply to people that email you, to let them know that their email will be moderated by others before you receive it, with an option to rescind or revise the original message? [Only the first time this person contacts you] Do you think this would create too much of a barrier for strangers wishing to have legitimate communication with you?

- How would you feel if you received such an automated email after sending an email to someone? Would you be ok with sending the email through?

- What kind of turnaround would you expect/want/need for receiving moderated emails? (1 hour, 12 hours, 24 hours, a week?)
- Moderators could tag emails and/or write explanations of their decision. What information would you want to get from moderators about their decisions? If explanation, maybe paraphrase the content of the email to you? If tags, what kind of attributes?
- Would you want moderators to build up a whitelist of people/terms or blacklist of people/terms for you, or would you want to do that yourself?
- How would you like to communicate with moderators?
- Would you ever want to contact them to alert them of something or have them contact you in certain cases? What kinds of cases? Different for friend moderators vs. strangers?
- Would you want to be able to see the emails that they rejected? If so, how? A separate folder in email, in your inbox but with a tag, or on a webpage?
- What do you want moderators to know about you (preferences, exceptions) before they do their work?
- How would you prefer emails that might need escalation (like to law enforcement) be dealt? Would you want the moderators to flag them and send them to you to be dealt with? Would you rather them deal with it?
- What constitutes an email that needs escalation?
- What actions, if any, would you want the moderators to ask of people that send harassing emails? They could: ignore the email and send it to the trash, email back the harasser to ask them to rephrase it, let the harasser know that the email was never sent to the recipient. Any others?
- How would you feel about us storing the harassing emails on our server, or sharing the content of these emails with us? For instance, this data could be

used to train models for you that detect emails that are likely to be harassment. Or, the emails could be publicized (with private info redacted) on our website as a deterrent.

**Training**

- What qualifications would you want to have in a paid moderator?
- What qualifications would you want to have in a volunteer (stranger) moderator?
- Would you want your moderators (friend, crowd, volunteer) to be trained at all? What kind of training do you have in mind?
- What about training them on your personal desires/needs for how emails should be handled?

# Bibliography


[1] CJ Adams and Lucas Dixon. Better discussions with imperfect models, September 2017.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Maeve Duggan. Online harassment 2017. The Pew Research Center, July 2017.

[8] Robert Faris, Amar Ashar, Urs Gasser, and Daisy Joo. Understanding harmful speech online. December 2016.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Randi Lee Harper. Good game auto blocker, 2014.

[14] Randi Lee Harper. Putting out the twitter trashfire. Art + Marketing, February 2016.

[15] Jacob Hoffman-Andrews. Blocktogether, 2017.

[16] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving google's perspective API built for detecting toxic comments. *CoRR*, abs/1702.08138, 2017.

[17] Google Jigsaw. Perspective api, 2017.

[18] 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.

[19] 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.

[20] 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.

[21] Amanda Lenhart, Michele Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. Online harassment, digital abuse, and cyberstalking in america. Data & Society, January 2017.

[22] Alice E Marwick and Ross W Miller. Online harassment, defamation, and hateful speech: A primer of the legal landscape. June 2014.

[23] J. Nathan Matias. High impact questions and opportunities for online harassment research and action, August 2016.

[24] 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.

[25] Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on twitter. *CoRR*, abs/1706.01206, 2017.

[26] 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.

[27] 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.

[28] Tamara Shepherd, Alison Harvey, Tim Jordan, Sam Srauy, and Kate Miltner. Histories of hating. *Social Media + Society*, 1(2), 2015.

[29] Aaron Smith and Maeve Duggan. Crossing the line: What counts as online harassment? The Pew Research Center, January 2018.

[30] Working to Halt Online Abuse. Whoa comparison statistics 2000-2013, 2013.

[31] 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.

[32] Charlie Warzel. "a honeypot for assholes": Inside twitter's 10-year failure to stop harassment, August 2016.

[33] Charlie Warzel. Twitter is still dismissing harassment reports and frustrating victims. Buzzfeed, July 2017.

[34] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. pages 78–84, 2017.

[35] 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.

[36] 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.

[37] 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.