Squadbox: A Tool To Combat Online Harassment Using Friendsourced Moderation

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

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

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

CHI'18 Extended Abstracts, April 21–26, 2018, Montreal, QC, Canada
ACM 978-1-4503-5621-3/18/04.
https://doi.org/10.1145/3170427.3186498

Abstract

Communication platforms have struggled to provide effective tools for people facing harassment online. Rather than relying on platforms, we consider how harassment recipients can harness their personal community for support. We present *Squadbox*, a tool to help recipients of email harassment coordinate a “squad” of friend moderators to shield and support them during attacks. Moderators intercept email from strangers and can reject, organize, and redirect emails as well as collaborate on filters. Harassment recipients can highly customize the tool, choosing what messages go through, how moderators should handle particular messages, and if and how they receive rejected messages.

Author Keywords

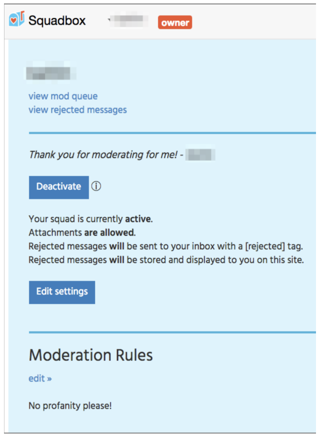
online harassment; email; moderation; private messages; friendsourcing; crowdsourcing; social media

ACM Classification Keywords

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

Introduction

According to a recent report by the Pew Research Center [3], nearly half of internet users in the United States have experienced some form of online harassment or abuse.



[redacted] [redacted]

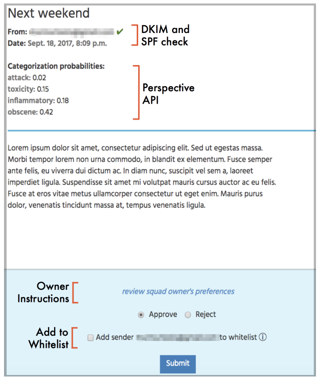
Thank you for moderating for me! - [redacted]

Your squad is currently **active.**
Attachments **are allowed.**
Rejected messages **will** be sent to your inbox with a [rejected] tag.
Rejected messages **will** be stored and displayed to you on this site.

Moderation Rules

No profanity please!

Figure 1: An owner's view of the information page for their squad.



Next weekend

From: [redacted] ✓
Date: Sept. 18, 2017, 8:09 p.m.

Categorization probabilities:
attack: 0.02
toxicity: 0.15
inflammatory: 0.18
obscene: 0.42

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed ut egestas massa. Morbi tempor lorem non urna commodo, in blandit ex elementum. Fusce semper ante felis, eu viverra dui dictum ac. In diam nunc, suscipit vel sem a, laoreet imperdiet ligula. Suspendisse sit amet mi volutpat mauris cursus auctor ac eu felis. Fusce at eros vitae metus ullamcorper consectetur ut eget enim. Mauris purus dolor, venenatis tincidunt massa at, tempus venenatis ligula.

● Approve ○ Reject

□ Add sender [redacted] to whitelist ⓘ

Figure 2: The moderator's view of a message. Squadbox provides information such as sender verification results and Perspective API scores along with the message.

Unfortunately, solutions for combating online harassment have not kept pace. Common solutions such as user blocking and word-based filters are blunt tools that cannot cover many forms of harassment, are labor-intensive for people suffering large-scale attacks, and can be circumvented by determined harassers. Even so, platforms have been criticized for their slow implementation of said features.

Researchers have built machine learning models to detect harassment [2, 6], but caution that such models should be used in tandem with human moderators, due to inaccuracy and bias. Given the inability of wholly automated systems to solve harassment, many recipients turn to friends [5] and community-based anti-harassment tools. These tools include Heartmob [1], which provides a volunteer support network, and collaborative blocklist tools like BlockTogether [4].

Building on this work, we present Squadbox [5], a tool that allows users to coordinate a "squad" of trusted individuals to moderate messages when they are under attack. Using our tool, the "owner" of a squad can automatically forward potentially harassing content to a moderation pipeline. When a message arrives, a moderator makes an assessment, adding annotations and rationale as needed. The message is then handled in a manner according to the owner's preference, such as having it delivered with a label, filed away, or discarded. Rather than making decisions for users about how exactly to use the tool, we designed it to be highly customizable to different possible owner-moderator relationships and usage patterns. At the same time, we aim to *scaffold* the owner and moderator actions so they can be performed more easily than current jerry-rigged approaches. Our initial implementation targets email, as email is particularly weak on anti-harassment tools and has a standard, powerful API. The system can be extended to any platform with a suitable API, and we plan to do so.

Squadbox: A Friendsourced Moderation Tool

We describe Squadbox[1] user scenarios for the workflows shown in Figure 3, followed by features and implementation.

User Scenarios

Flow A: Squadbox as a public contact address. Adam is a journalist, harassed on Twitter about his articles. He wants a publicly-listed email address to receive tips from strangers, but is hesitant, fearing harassment. Adam creates a Squadbox address, `adam@squadbox.org`. He enlists two coworkers to be moderators because they know him and his work well. Adam uses `adam@squadbox.org` as a public email address. Any email sent there goes through his squad first. Now he can open himself up to the public without risking further harassment.

Flow B: Squadbox with an existing email account. Eve is a professor. She has a publicly-listed email address through the university. 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, but can't communicate with her collaborators when she's under attack. Eve sets up a squad and asks her spouse and a friend to serve as moderators. She sets up a whitelist and filters so that only strangers' emails go to Squadbox. She can turn on Squadbox when she starts getting harassment, but then turn it off when it dies down. A second scenario for Flow B involves Julie, dealing with harassment from an ex-significant other. She can't 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 emails from her harasser to her squad. Her moderator separates out and returns information about coordination, while redacting harassing content.

[1] http://squadbox.org/

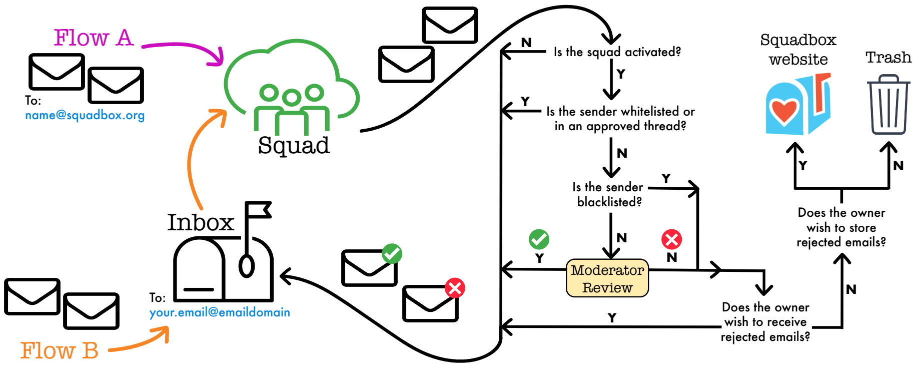

Flow A
To:
name@squadbox.org
Squad
Inbox
To:
your.email@emaildomain
Flow B
Is the squad activated?
N
Y
Is the sender whitelisted or in an approved thread?
Y
N
Is the sender blacklisted?
Y
N
Moderator Review
Y
N
Does the owner wish to receive rejected emails?
Y
N
Does the owner wish to store rejected emails?
Y
N
Squadbox website
Trash

Figure 3: Diagram of the flow of emails through Squadbox, including Flow A, which allows users to have a public moderated account, and Flow B, which allows users to get their current account moderated. From there, various settings define whether emails get moderated and where they go.

Features for Giving Owners Customization Capabilities
Messages can have **tags** applied to them to give information to the intended recipient without requiring them to view the message. Rejected messages automatically have the "rejected" tag applied; the moderation interface allows adding more tags indicating common reasons why a message might be rejected, such as “insult” or “profanity”. Owners can choose whether they want rejected messages delivered to their inbox with tags in the subject line, which enables them to add filters in their mail client customizing where those messages go. They can also have rejected messages stored on the Squadbox website, where they can be grouped and sorted by tag.

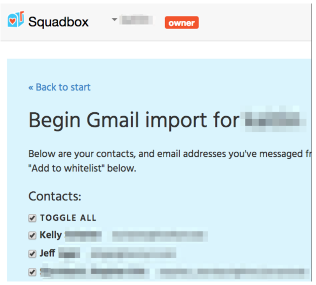
As owners differ on definitions of harassment and desired moderator actions, they give **custom instructions** via a freeform text box. As owners might want to know moderators’ rationale for rejecting particular messages, moderators can provide an **explanation** for their decision or a **summary**, to be shown to the owner along with the message.

Features for Reducing Moderator Load and Increasing Privacy
Squadbox supports **filtering** by sender whitelists and blacklists, meaning emails from specified senders will be automatically approved or rejected, respectively. Such filters partially alleviate concerns about slow moderation turnaround time, and give owners more control over what their moderators see. There is significant room to expand this 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.

Owners can set Squadbox to **automatically approve replies** to threads where the initial post is approved, with the ability to opt back in to moderation as needed. This provides fine-grained control over what moderators see, reduces the number of messages to review, and makes extended conversations less hindered by delays.

To better accommodate fluctuating volumes of harassment, owners can **deactivate** their squad so that all messages are auto-approved. When the squad is reactivated, previously defined settings and filters take effect.

Features for Reducing Secondary Trauma to Moderators
To prevent disruption, we give moderators **control over viewing harassment**, only showing them messages when they choose to visit the website. When a message arrives we notify the least recently notified moderator, and only if they haven’t been notified in at least 24 hours. This makes it easier for moderators to step back and **limit their work**. We plan to enable moderators to temporarily take a break from the system, and to set hard limits on moderation time/volume. We also plan to provide training and support resources.



« Back to start

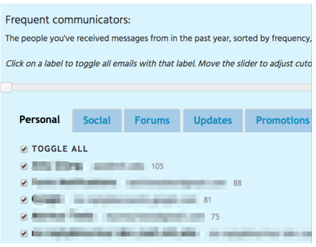
Begin Gmail import for [redacted]

Below are your contacts, and email addresses you've messaged fr
"Add to whitelist" below.

Contacts:

☑ TOGGLE ALL
☑ Kelly [redacted]
☑ Jeff [redacted]
☑ [redacted]

Figure 4: Gmail contacts are imported to generate whitelist suggestions.


Frequent communicators:

The people you've received messages from in the past year, sorted by frequency,

Click on a label to toggle all emails with that label. Move the slider to adjust cuto

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

Figure 5: Contacts are grouped by category and sorted by communication frequency.

Features for Giving Moderators Context and Information
We show the entire thread of messages to a moderator to provide **context**. 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. As shown in Figure 2, we display whether the message passes **sender verification** via SPF and DKIM checking, which cryptographically detect *spoofing*—senders pretending to be other senders. We also show **machine-generated scores** of messages' likelihood of being harassment, generated by the Perspective API[2].

System Implementation
Squadbox is a Django web application. Data is stored in MySQL, and attachments in Amazon S3. It interfaces with a Postfix SMTP server using the Python Lamson library. **Flow A** works just like a moderated mailing list with one member. **Flow B** 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 sets filters on their email client to forward unmoderated messages to Squadbox and to keep already-moderated messages. For Gmail users, we leverage the API to generate whitelist suggestions, as in Figure 4, and to programmatically create filters.

Demo Goals and Conclusion

Our goal in demonstrating Squadbox is to receive feedback on current and planned features. We'll have interactive demos of the moderation interface and Perspective API, showing attendees how the system works and how AI techniques can assist, but not replace, human moderators. We are actively working to get Squadbox ready for a public release, and hope that demoing will help us find potential users of the system and contributors to the project.

[2] www.perspectiveapi.com

REFERENCES

1. Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 24 (Dec. 2017), 19 pages.
2. Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3175–3187.
3. Maeve Duggan. 2017. Online Harassment 2017. The Pew Research Center. (11 July 2017). Retrieved September 8, 2017 from `http://www.pewinternet.org/2017/07/11/online-harassment-2017/`
4. Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Transactions on Computer-Human Interaction* 25, 2, Article 1 (March 2018), 33 pages.
5. Kaitlin Mahar, Amy X. Zhang, and David Karger. 2018. Squadbox: A Tool To Combat Online Harassment Using Friendsourced Moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA.
6. Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1391–1399.