Combating Online Harassment with Friendsourcing

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

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

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

Abstract
Current approaches to combating online harassment put responsibility in the hands of the person being harassed to report, filter, or block harassment, or the hands of platforms to ban bad actors. However, in many cases, harassment recipients are too overwhelmed to effectively combat their harassment alone, and platform-wide bans of an individual can oftentimes be too blunt of a solution. To complement the above approaches, we argue that there should be more tools for harassment recipients to invoke the help of friends or their community to combat harassment and respond to harassers.

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
Online harassment has been a longstanding problem on communication platforms without clear or comprehensive solutions to date. The issue is complex in that harassment can mean different things to different people, and comes in many forms [5]. First, people have very different defini-

tions for what constitutes online harassment [6]. In terms of content, some consider uncivil language as harassing, while others do not. Some find deeply personal or graphic material a boundary violation, while others do not. In many cases, the definition can change based on context, such as the identity of the harasser or the recipient. Sometimes, messages can have innocuous content but still be considered harassing due to the sender's persistent activity.

Another issue is that people have very different ways that they want to respond to harassment. Some people seek to have their harassers banned from platforms, or attempt to block harassers from reaching them. Others prefer to reach out to harassers to try to get them to change their behavior. Still others choose to publicize their harassment, to bring awareness to the problem or to reveal their harassers' identities. Again, this depends on the context.

Given such wide variation, a single, universal solution or set of guidelines will leave out many people. For example, a blunt tool such as a platform-wide ban can be effective only in certain circumstances, and should not be the only recourse that harassment recipients can employ. When wielded broadly or indiscriminately, platform-wide bans place enormous power in the hands of platforms to decide what is permissible. Indeed, many users have complained about the lack of transparency oversight into platform decisions and their processes. On the other hand, if used only in narrow, clear-cut cases of policy violation, platform-wide bans will miss many cases of harassment that still cause significant harm to recipients [3].

Instead of limiting harassment recipients to solutions that depend on the whims of platforms and that focus on reducing harassers' *ability to speak* on a platform, we can in addition consider tools to increase harassment recipients' *ability to decide what to hear*. This also allows each recipient to take the personal approach that they prefer to counter bad actors.

However, this also does not mean that we should put everything on the shoulders of harassment recipients. While recipients best know the details of their harassment and their desires for responding to it, the task of dealing with harassment can be overwhelming and emotionally draining. Existing tools to help recipients with this task are insufficient. Sometimes a recipient cannot or does not want to block a harasser, such as when they still need to communicate with the harasser for other reasons, or if they need to keep aware of possible threats. Word-based filters have many false positives and false negatives that need to be sorted through and require constant upkeep as language changes. Both of these can be circumvented by harassers with little effort. The diversity in user preferences and the contextual nature of harassment also makes other solutions such as a uniform learned model for detecting harassment difficult. Finally, responding is not only emotionally taxing and repetitive, but can lead a recipient to open themselves up to even more harassment by provoking the harasser.

Given that *platform-wide* solutions can be too blunt and *individual* solutions can be too taxing, we instead propose a suite of *friendsourced* techniques, or techniques empowering the friends or community of a harassment recipient to assist or intervene. There are several reasons why such strategies are likely to be successful and should exist alongside platform-wide and individual tools for when those tools fall short. First, techniques of this kind invoke the advantage that harassment recipients may sometimes have, which is strength in numbers. The people who are motivated to help recipients often outnumber those who are targeting someone, yet due to the design of current systems, highly motivated individual actors can have outsize impact on an

individual or community. Second, some existing tools such as volunteer support networks such as HeartMob [1] and shared Twitter blocklists such as BlockTogether [4] already make use of a collaborative strategy to combat harassment. These approaches suggest more avenues for support, as well as potential platform and tool designs to support such techniques. Finally, many harassment recipients are already using such strategies to mitigate harassment, but in a way that circumvents platform designs, such as by giving friends their passwords to delete messages [5].

Friendsourced Strategies to Combat Harassment

We present four possible ways for friends and communities to help recipients of harassment. These are informed by interviews we have conducted with 18 people who have experienced harassment across many different platforms [5].

Moderation of Harassing Messages

There is a great deal that friends can do to prevent harassment from disrupting someone's day-to-day. As previously mentioned, some recipients already use strategies such as giving friends their password, or forwarding emails unopened for friends to check. Our tool Squadbox [5] makes this process easier and more privacy-preserving, by allowing recipients to set up filters that determine what messages friends should moderate. There are many possibilities for complex filters that allow harassment recipients and moderators to enhance their privacy or reduce work, respectively. In addition to simply moderating content, moderators can help with managing blocklists and word-based filters to make their own work more automated over time. Similarly, personalized machine learning models could be trained on individual moderator inputs to suggest filters, score and sort content, and otherwise help moderators prioritize their time [2]. Thus instead of recipients managing all this alone, multiple people can take part to make the task less overwhelming. Additionally, much like shared Twitter blocklists, some of this work can be shared with other harassment recipients who encounter similar harassment.

Documenting and Reporting of Harassment

If harassment recipients wish to have platforms or law enforcement informed about their harassment, they currently must go through a lengthy process of collecting and documenting their harassing messages. Tracking the status of reports or keeping track of a particular harasser over time and across platforms also requires sifting through many messages. This is another area where friends can help, particularly if they are already moderating messages. In addition to moderation, they can flag or tag messages into certain categories, or use automated tools to take a snapshot of the message or capture context to analyze later. More work in concert with legal scholars is needed to determine what apparatus is necessary to create admissible evidence.

Respond to Harassers

Many harassment recipients that we spoke with were divided on whether and how to respond to harassers. Some recipients felt that it would be a waste of time or simply lead to more harassment. Other recipients mentioned times where they had responded and harassers apologized or changed their tone or stance. Recipients that were interested in having friends respond to harassers were still concerned about the safety of their friends, as well as friends overstepping or even harassing the harasser. One idea brought up by a recipient was to create template responses to different types of harassment that moderators could send in the recipient's name. There are other cases where bystander intervention could take a larger role, for instance in public settings where it may be important to signal to the

harasser or other readers that the behavior violates a community norm.

Support Recipients of Harassers
Finally, harassment recipients have talked to us about how important it was that they received words of encouragement from friends and their community when they were undergoing harassment. Friends and community members could combat the emotional toll of attacks by expressing their support and appreciation to the harassment recipient, much like how HeartMob operates [1].

A Suite of Anti-Harassment Tools

Because current approaches to combating harassment such as platform bans or personal blocking or filtering have many downsides, tools that enable friends and community members to assist recipients of harassment could fill an important gap. The tools we propose could sit alongside existing tools to together handle a broad range of harassment scenarios. When harassment is a minor or infrequent issue, tools for an individual may be enough, as was the case for many harassment recipients we spoke to when between short spikes of heavy harassment. When harassment is not directed to an individual but instead is speaking about that individual in a public space, such as with “doxing" and revenge porn, strategies such as a platform-wide ban may be necessary.

Where Platforms Fit In

Who should be responsible for building and integrating such anti-harassment tools? On the one hand, having cross-platform solutions would mean fewer interfaces to deal with and would allow users to have a single location for moderators, preferences, and storing or interacting with their harassment. It would also make it easier to document cases of harassment on multiple platforms at once or follow an individual who is using multiple platforms to harass. On the other hand, friendsourced moderation tools may require access beyond what public APIs provide in order to effectively function, meaning that only the platform itself could implement the desired functionality.

Our Experiences with Online Communities

The authors of this paper have been speaking with harassment recipients for the last year and a half to learn about their experiences while developing the Squadbox tool. We have presented the Squadbox tool at MozFest 2017 and have received feedback from organizations involved with anti-harassment initiatives, such as Hollaback!, OnlineSOS, and Jigsaw, as well as prominent harassment recipients and anti-harassment activists. The Squadbox project was kickstarted at the Beyond Comments workshop hosted by the Coral Project at the MIT Media Lab and grew out of the Murmur project [7], a related system for re-imagining the mailing list. Besides this work, the authors have been active in conducting research and building tools to allow end users to better manage their group communication experiences.

In terms of personal experiences, all authors are active on online community and social media platforms. The first author got her start participating in online communities via CreateBlog and Neopets forums and Xanga blogging in middle and high school, graduating to writing on Blogger and lurking on Reddit in college, and now spends a great deal of time on Twitter. The second author was very active on Tumblr for most of high school, and now spends much of her time online on Facebook groups and subreddits devoted to animals, memes, and skincare.

REFERENCES

1. 2017. Heartmob. (2017). Retrieved September 8, 2017 from `https://www.iheartmob.org`

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

3. 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.

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

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

7. Amy X Zhang, Mark S Ackerman, and David R Karger. 2015. Mailing Lists: Why Are They Still Here, What's Wrong With Them, and How Can We Fix Them?. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 4009–4018.