Online Harassment and Content Moderation: The Case of Blocklists

SHAGUN JHAVER, Georgia Institute of Technology
SUCHETA GHOSHAL, Georgia Institute of Technology
AMY BRUCKMAN, Georgia Institute of Technology
ERIC GILBERT, Georgia Institute of Technology

Online harassment is a complex and growing problem. On Twitter, one mechanism people use to avoid harassment is the *blocklist*, a list of accounts that are preemptively blocked from interacting with a subscriber. In this paper, we present a rich description of Twitter blocklists - why they are needed, how they work, and their strengths and weaknesses in practice. Next, we use blocklists to interrogate online harassment - the forms it takes, as well as tactics used by harassers. Specifically, we interviewed both people who use blocklists to protect themselves, and people who are blocked by blocklists. We find that users are not adequately protected from harassment, and at the same time, many people feel they are blocked unnecessarily and unfairly. Moreover, we find that not all users agree on what constitutes harassment. Based on our findings, we propose design interventions for social network sites with the aim of protecting people from harassment, while preserving freedom of speech.

CCS Concepts: •**Human-centered computing →Empirical studies in collaborative and social computing;** *Ethnographic studies;*

Additional Key Words and Phrases: Online harassment, moderation, blocking mechanisms, GamerGate, blocklists

ACM Reference format:
Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2, Article 1 (March 2018), 33 pages.
DOI: 0000001.0000001

1 INTRODUCTION

1.1 Online Harassment

In mid 2016, 25-year-old Erin Schrode was in the middle of her congressional campaign. She was aiming to become the youngest woman ever elected to the U.S. House of Representatives. Days before the election, she began receiving hate-filled emails and tweets from anonymous individuals who targeted her for being Jewish. One email said, "Get to Israel to where you belong. That or the oven. Take your pick" [2]. Another said, "... all would laugh with glee as they gang raped her and then bashed her bagel-eating brains in" [38]. On election day, Schrode switched on her

This work is supported by ...

Author's addresses: S. Jhaver, S. Ghoshal, E. Gilbert and A. Bruckman, School of Interactive Computing, Georgia Institute of Technology
Contact author: Shagun Jhaver.

DOI: 0000001.0000001

computer and found that her campaign website had been hacked and all references to her name were converted to Adolf Hitler. Over the next few months, the attacks grew more numerous and repulsive. Some posters attached doctored photos of Schrode in their messages - one sent a photo of her wearing a Nazi style yellow star; another sent an image of her face stretched onto a lampshade. Every time Schrode looked at any of her social media feeds or emails, she was reminded that she was unwelcome and told that she was inferior. She often felt lonely and suffocated. "You read about these things in the news," she said, "but it's so unreal when it targets you" [2].

Schrode's experience is far from unique. In recent years, online harassment has emerged as a growing and significant social problem. According to a 2017 Pew Research study, 41% of American adults have experienced online harassment and 66% of adults have witnessed at least one harassing behavior online [16]. This study also found that social media is the most common venue in which online harassment takes place [16]. Many online offenders have turned social media platforms into forums to bully and exploit other users, threaten to harm or kill them, or reveal sensitive information about them online.

The problem of online harassment is particularly prevalent on Twitter[1] [20, 31]. Some critics have worried that Twitter has become a primary destination for many trolls, racists, misogynists, neo-Nazis and hate groups [53]. Twitter has indeed found itself ill-equipped to handle the problem of online harassment, and its CEO has declared, "We suck at dealing with abuse and trolls on the platform and we've sucked at it for years" [45].

In this article, we use Twitter blocklists, a third party blocking mechanism aimed at addressing online abuse on Twitter, as a vehicle to explore the problem of online harassment. We review different experiences and perceptions of online harassment. We find that many Twitter users feel that existing moderation tools on Twitter fail to provide them with adequate protection from online abuse, and they circumvent the limitations of such tools by developing, deploying and promoting third-party moderation tools like blocklists. We investigate how the use of blocklists is perceived by those who use them and by those who are blocked because of them.

1.2 Blocking on Twitter

In this section, we explain the use of blocking and muting mechanisms on Twitter. Following this, we describe Twitter blocklists.

Many platforms have implemented moderation mechanisms to discourage antisocial behavior such as trolling and harassment. These mechanisms include using a small number of human moderators who manually remove abusive posts [25], moderating through a voting mechanism where registered users up-vote or down-vote each submission [27], and allowing users to flag abusive content [11]. Another mechanism that many platforms primarily rely on is providing users the ability to mute, block, or report offensive users (Figure 1).

On most platforms, and particularly on Twitter, blocking or muting an account allows a user to stop receiving notifications from that account, and that account's posts don't appear on the user's timeline or newsfeed [47]. The difference between blocking and muting is as follows: blocking an account prevents that account from viewing the blocker's posts or sending direct messages to the blocker. In contrast, a muted account can still view the user's posts, "favorite" them, and reply to them. Muting an account is more socially delicate than blocking it: a muted user is not notified that he is muted, and he may continue posting to the user who muted him without realizing the

[1] https://twitter.com. Founded in 2006, Twitter is a microblogging platform that allows its users to post 140-character messages, called tweets, about any topic, and follow other users to read their tweets. Recent news articles suggest that Twitter is one of the five most popular social network sites worldwide [33]. As of the first quarter of 2017, it averaged at 328 million monthly active users [44].

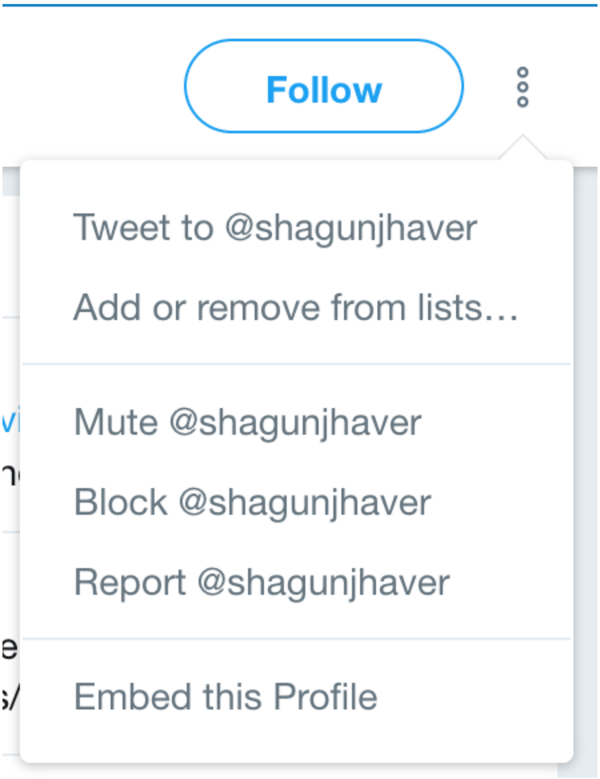


Fig. 1. Twitter provides options to mute, block, and report offensive users.

receiver cannot see his posts, whereas a blocked user immediately realizes that he is blocked if he attempts to post to the blocker. If the blocked user accesses the blocker's profile, he sees the following message:

> "You are blocked from following @[blocker] and viewing @[blocker]'s Tweets. Learn more"

Blocklists are third-party Twitter applications that extend the basic functionality of individual blocking by allowing users to quickly block all accounts on a community-curated or algorithmically generated list of block-worthy accounts [20]. Perhaps the most popular type of blocklists are anti-harassment blocklists that aim to block online harassers en masse. The use of decentralized moderation mechanisms like anti-harassment blocklists takes some pressure off the central Twitter moderators so that they don't have to be as strict in their moderation. Everyone has slightly different boundaries, and the use of these lists can provide users an experience that is more customized to their needs [20]. However, not everyone who is put on anti-harassment blocklists sees himself as a harasser. Some of the users blocked by these lists may think of themselves as perfectly reasonable individuals. We will expand on this problem and other limitations of blocklists in our findings.

Next, we discuss two different Twitter applications that serve as blocklists.

1.2.1 Block Bot. Block Bot[2] was the first blocklist implemented on Twitter. It is a socially curated blocklist where a small group of moderators coordinate with one another and make complex decisions about which Twitter users to put on a shared list of blocked accounts.

Block Bot emerged out of the use of hashtag #BlockSaturday[3] on Twitter. In 2012, a Twitter user began identifying accounts that he felt were worthy of blocking and started posting tweets containing the usernames of these accounts along with the #BlockSaturday hashtag. This was done so that his followers and anyone following the hashtag #BlockSaturday could block those accounts [6]. As more users began posting such tweets and this trend became more popular, a few users expressed a need to automate the process of blocking. This led to the creation of Block Bot. Its developers made creative use of Twitter APIs that were developed to support third-party clients such as smartphone applications [20]. The use of this blocklist allowed users to collectively curate lists of Twitter accounts that they identified as harassers and block them together quickly and easily.

In its initial days, Block Bot was primarily used to serve the atheist feminist community and block individuals who opposed the rise of the Atheism+[4] movement. Block Bot later expanded its goals to block supporters of GamerGate movement [18] , users who harass transgender people, and other abusive accounts. This blocklist allows moderators to sort blocked users into three categories of offensiveness – nasty, unpleasant and annoying, and allows subscribers to pick the level of offensiveness they would like to excise from their Twitter feeds [22].

1.2.2 Block Together. Block Together[5] is a web application that serves as a "centralized clearing-house" for many blocklist curators and subscribers [20]. Like Block Bot, Block Together is a Twitter application that was developed by volunteers to combat harassment on Twitter. It was released by third-party software developers at the Electronic Frontier Foundation [15]. In contrast to Block Bot, which hosts a unique list of blocked accounts, this application hosts many different lists of blocked accounts. Block Together allows Twitter users to share their own list of blocked accounts that other users can subscribe to (Figure 2). It also gives the subscribers an option to block accounts that are newly created or have fewer than 15 followers. This helps combat trolls who create new accounts on Twitter after they find themselves being blocked. Although Block Together was created to address online abuse, it now hosts many blocklists that serve vastly different purposes, including those that block spam accounts and those that block ISIS critics. However, in this article, we restrict our discussion to anti-abuse blocklists.

Next, we discuss Good Game Auto Blocker, a popular blocklist that is hosted by Block Together.

GamerGate and Good Game Auto Blocker

Online harassment often occurs as a result of coordinated harassment campaigns organized by hate groups that overwhelm a target by synchronously flooding his or her social media feeds [20]. One group that has recently gained attention in the popular media for coordinating such campaigns is GamerGate[6]. Although conversations about GamerGate can be found on many online sites like Reddit [24], Voat, 8chan and YouTube, Twitter has emerged as one of the most popular sites for

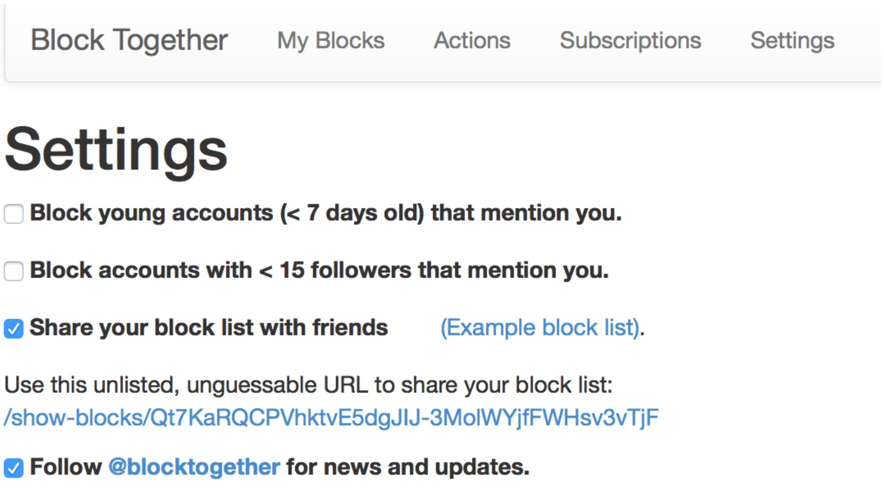
[2] http://www.theblockbot.com

[3] #BlockSaturday is a wordplay on a popular trend of using hashtag #FollowFriday. Twitter users used #FollowFriday in their posts to recommend to their friends on Twitter other handles to follow.

[4] Atheism+ is a movement that originated in August 2012 by blogger Jen McCreight. It encouraged progressive atheists to cater to issues other than religion, such as social justice, feminism, racism and homophobia [39].

[5] https://blocktogether.org

[6] GamerGate is an online social movement that emerged in response to a series of controversial events surrounding game developer Zoe Quinn [23]. The supporters of the movement insist that GamerGate stands for ethics in gaming journalism. However, a number of media articles portrayed the movement as a hate group, and claimed that users supporting the



Settings

□ Block young accounts (< 7 days old) that mention you.

□ Block accounts with < 15 followers that mention you.

☑ Share your block list with friends (Example block list).

Use this unlisted, unguessable URL to share your block list:
/show-blocks/Qt7KaRQCPVhktvE5dgJIJ-3MolWYjfFWHsv3vTjF

☑ Follow @blocktogether for news and updates.

Fig. 2. BlockTogether settings for blocking accounts on Twitter.

discussing this movement. Many Twitter users who oppose GamerGate consider its supporters as harassers and feel a need to block them en masse. This led to the creation of Good Game Auto Blocker (GGAB), a blocklist aimed at blocking GamerGate supporters on Twitter.

GGAB uses Block Together to implement blocking on Twitter. The current procedure for blocking users on GGAB is unknown, but at least in its initial days, GGAB used a predominantly algorithmic approach to curate the list of block worthy accounts [19]. It collected the followers for a short list of prominent #GamerGate contributors on Twitter. If anyone was found to be following more than two of these supporters, they were added to the list and blocked. This blocklist also used a periodically updated white-list of users who satisfied this criterion but were false positives. The accounts on the white-list were unblocked. For this blocklist, a block removal can be appealed by providing a link to one's Twitter profile and an explanation of why an exception should be made.

In this study, we interviewed users who subscribed to GGAB or Block Bot blocklist as well as those who are blocked on such lists in order to understand the motivations, benefits and limitations of the use of this novel socio-technical, anti-abuse mechanism. Our findings indicate that many users find blocklists to be quite effective in addressing online harassment. However, the widespread use of blocklists, as they are currently implemented, can also lead to many problems. For example, we discovered the problem of a "blocking contagion": when a popular blocklist is forked to create multiple other lists, false positive accounts on the original blocklist end up getting blocked by users who subscribe to any of the several forked lists. This results in a large number of users inadvertently censoring these false positive accounts. We discuss this phenomenon and other shortcomings of blocklists in our findings. We also discuss the advantages and limitations of using GGAB over using Block Bot.

1.3 Research Questions

We explore the following research questions in this study:

movement engage in death threats, rape threats and doxing among other harassment activities. Mortensen provides a detailed account of the progression of events that helped GamerGate gain popular attention [34] .

1) How do perceptions of online harassment vary between users who subscribe to blocklists and users who are blocked on these lists? What behavior patterns do blocklist subscribers identify as instances of online harassment?
2) What motivates users to subscribe to anti-abuse blocklists and how do their experiences change after subscribing? What are the advantages and challenges of curating blocklists using human moderators? How do blocked users perceive the use of anti-abuse blocklists?

1.4 Contributions

In this paper, we contribute, first, a rich description of Twitter blocklists - why they are needed, how they work, and their strengths and weaknesses in practice. Second, we contribute a detailed characterization of the problem of harassment. We include the perspective of people accused of harassment, which is often omitted from discussions of this topic. As we will see, both those who suffer harassment and those who are accused of it are diverse groups. This diversity matters when we plan solutions. Further, we explore the idea that the flip side of harassment is understanding across differences - these problems are intertwined. Finally, we contribute a set of design challenges for HCI in addressing these issues.

The remainder of this article is organized as follows: we start by discussing related work. Next, we present our methods of data collection and analysis. We discuss the qualitative analysis of our interviews and observations, focusing on our findings on online harassment and Twitter blocklists. In the final section of the paper, we build on our findings to suggest a broad set of design opportunities that can help address online harassment on social media. We close with possible future directions of this study.

2 RELATED WORK

2.1 Online Harassment

Like face-to-face harassment, online harassment has a deeply negative impact on its recipients. They may experience significant emotional problems such as anxiety and depression and, in extreme situations, may even commit suicide [1]. A 2017 Pew Research study found that 13% of US adults experienced mental or emotional stress as a result of online harassment, another 8% of adults indicated that they had problems with friends or family because of online harassment, and 7% said that these experiences caused damage to their reputation [16]. According to a recent Data & Society Research Institute report, 27% of internet users self-censor their online postings out of fear of online harassment [28]. Furthermore, 43% of recipients of online abuse had to change their contact information to escape their abuse [28].

Young people, minorities and women are particularly vulnerable to such abuse [15, 16, 28]. In a national study of middle and high school students, 60 percent of lesbian, gay, bisexual, and transgender (LGBT) youth reported being harassed based on their sexual identity and 56 percent of them felt depressed as a result of being cyberbullied [10].

While a considerable body of research has been conducted on face-to-face harassment, less is currently known about online harassment. One of the primary challenges of addressing online harassment is that it is often in dispute what online harassment actually means. Lenhart et al. argue that “online harassment is defined less by the specific behavior than its intended effect on and the way it is experienced by its target” [28]. Many users who are accused of perpetrating harassment complain that simple disagreement on their part is often portrayed as harassment by other users. In their study of users accused of online harassment, Jhaver et al. found that “individuals often have more complex views than stereotypes predict,” and it is important to differentiate sincere misunderstanding from deliberate attacks [24]. They argue that although everyone would agree that posts of death and rape threats are abusive and should be regulated, it can be difficult to draw the line between passionate disagreement and deliberate harassment beyond such posts.

Danah boyd has identified four properties of SNSs that fundamentally alter the social dynamics of harassment and amplify its effects online: persistence, searchability, replicability, and invisible audiences [7]. The internet exacerbates the harassed users' injuries by extending the life of destructive posts. Search engines index content on the web, and harassers can put indexed abusive posts to malicious use years after they first appear [9]. Furthermore, by increasing the visibility of content, SNSs enable harassers to call for others to engage in abusive behaviors [51].

In this paper, we build on this prior work to understand more comprehensively and rigorously the different aspects of online harassment, its types, and its impact on those who are harassed.

2.1.1 Online Harassment on Twitter. Since its early days, Twitter has positioned itself as a platform for free speech. This supported the rapid gain in popularity of Twitter, and helped it play a critical role in many recent social movements – from the Arab Spring to Black Lives Matter [12, 29]. However, this maximalist approach to free speech also created conditions for online abuse on the platform [53]. A Buzzfeed report on Twitter noted that the platform treated online abuse as a "perpetual secondary internal priority" and allowed it to grow as a chronic problem over the last 10 years [53]. Twitter's unique design allows users who don't "follow" each other to interact, but it also makes it difficult to moderate content, because anyone can respond to any comment. This exacerbates the problem of abuse on the platform.

2.1.2 Free Speech and Online Harassment. Many American users cite First Amendment protections of the United States to argue that any regulation of online communities should be avoided, because it chills online expression. However, as Danielle Citron suggests, "there are speech interests beyond that of the harassers to consider" [9]. Online harassment often results in silencing of harassed users and therefore impinges upon their freedom of expression. Besides, free speech values are nuanced and they do not work as absolutes [9]. Certain categories of speech, such as true threats and defamation, do not enjoy constitutional protection [9]. Many critics also note that First Amendment protects speech from government censure but private individuals should be able to opt not to read or hear specific content.

Many participants leave online communities if they become too toxic. Some websites also shut down their comment systems when they are unable to cope with trolls [43]. Therefore, controlling online abuse is critical to maintaining the usability of online spaces.

2.2 Twitter Blocklists

Since the early days of the internet, social media sites have used blocking mechanisms to allow their users to filter the content they consume. Judith Donath describes how Usenet employed "killfiles," filters that allowed Usenet users to skip the unwanted postings [14]. If a user put someone in their killfile, he stopped seeing any more of their postings. The use of killfiles was found to be effective in keeping the newsgroups readable. However, Donath also describes the resentment of users blocked on Usenet:

"To the person who has been killfiled, Usenet becomes a corridor of frustratingly shut doors: one can shout, but cannot be heard" [14].

Donath characterizes killfiles as "a good example of a social action that is poorly supported by the existing technology" [14]. We assess in this article how far the technology has progressed to support the needs of social media users to ignore offenders by evaluating the use of contemporary blocking solutions, particularly blocklists on Twitter.

Stuart Geiger conducted a theoretical analysis of blocklists, and concluded that blocklists provide a concrete alternative to the default affordances of Twitter by facilitating a "bottom-up, decentralized, community-driven approach" for addressing online harassment [20]. Different individuals can have different perspectives on what online harassment entails, and where to draw the boundary between freedom of expression on the internet and online abuse. Geiger found that instead of a fixed Twitter-directed technological solution for addressing harassment, block-bots provide a social solution by allowing users with similar values to come together and engage in collective sensemaking. Geiger also notes that blocklists are "impactful in that they have provided a catalyst for the development of anti-harassment communities. These groups bring visibility to the issue and develop their own ideas about what kind of a networked public Twitter ought to be" [20].

This article contributes to updating Geiger's findings on blocklists [20]. We use empirical research methods to extend Geiger's work by incorporating arguments made by users affected by blocklists.

3 METHODS

Our IRB-approved study adopts a mixed methods approach. We use the results of a network analysis on Twitter to select our sample of participants for semi-structured interviews. We focus on Good Game Auto Blocker (GGAB), a popular blocklist currently in use. To understand the motivations and experiences of blocklist users, we interviewed 14 users who subscribe to GGAB and triangulated our findings by interviewing 14 users who were blocked on GGAB and analyzing our participants' posts on Twitter.

3.1 Participant Sampling

Sharan B. Merriam writes that "since qualitative inquiry seeks to understand the meaning of a phenomenon from the perspectives of the participants, it is important to select a sample from which the most can be learned" [32]. In selecting participants for this study, we used a purposive sampling approach. This approach advocates selecting participants who have rich information about issues of central importance to the research [32]. Inspired by Veldon and Legoze's study that combines network analytic approach with ethnographic field studies [50], we constructed a network of relevant users on Twitter and sampled the users most central to this network to recruit for interviewing. We expect that our centrality-based method of selecting interview participants helped us sample the dominant viewpoints of users.

As mentioned earlier, we interviewed two separate groups of participants: the first group is composed of users who were blocked on GGAB (hereafter referred to as UOB), and the second group contains people who subscribed to GGAB (hereafter referred to as SB). Next, we describe the details of how we sampled the participants for these two groups.

3.1.1 Selecting UOB Participants. We began by collecting the list of 9823 Twitter accounts blocked by GGAB. This list is publicly available on the BlockTogether website[7]. Next, we used the Twitter API[8] to retrieve the following information about each of the accounts on this list: (1) number of followers; (2) number of tweets issued; (3) date of account creation; (4) most recent tweet; (5) location; and (6) whether the account is verified.

We filtered out the accounts that were inactive (most recent tweet more than six months ago), created less than a year ago, verified (these are often accounts of brands and celebrities), located

[7] The list of accounts blocked by GGAB is available at http://tinyurl.com/ggautoblocker.

[8] https://dev.twitter.com/rest/reference/get/users/show

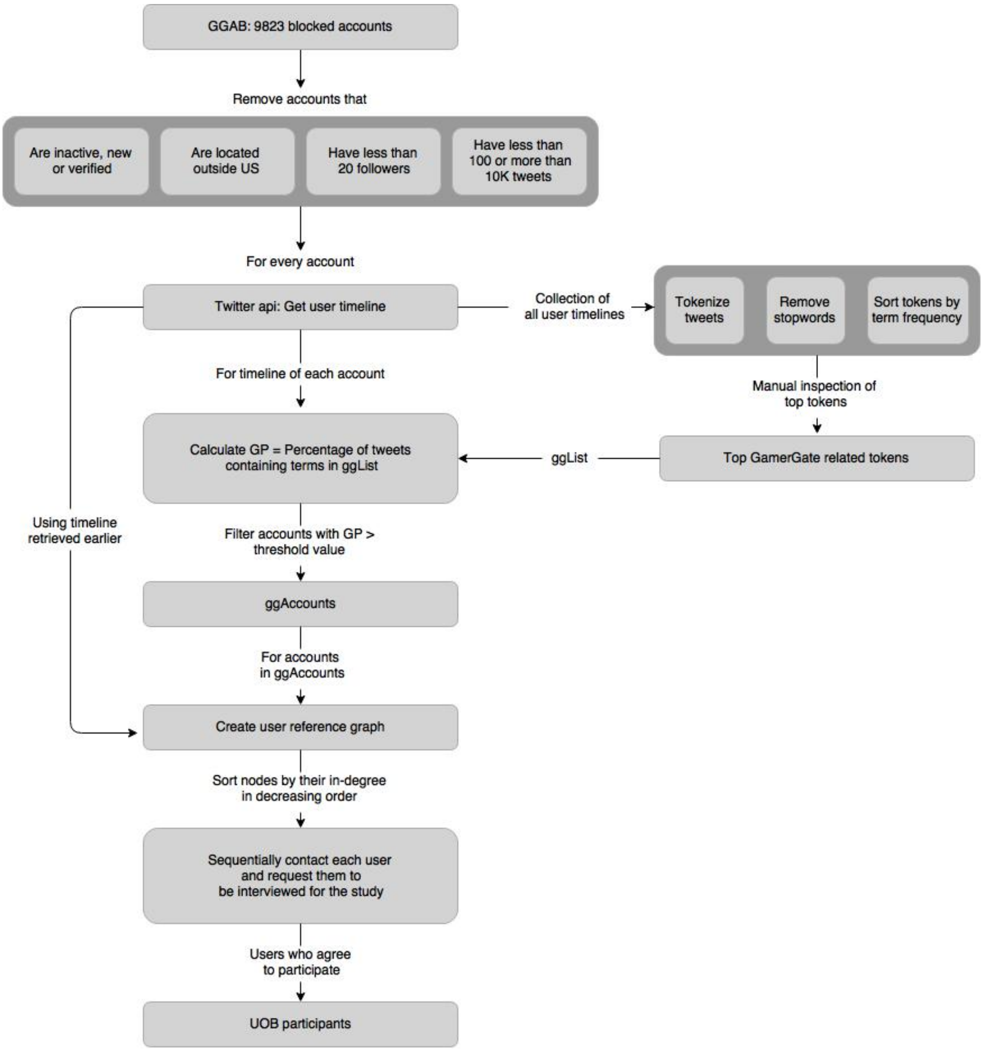

GGAB: 9823 blocked accounts
Remove accounts that
Are inactive, new or verified
Are located outside US
Have less than 20 followers
Have less than 100 or more than 10K tweets
For every account
Twitter api: Get user timeline
Collection of all user timelines
Tokenize tweets
Remove stopwords
Sort tokens by term frequency
Manual inspection of top tokens
Top GamerGate related tokens
ggList
For timeline of each account
Calculate GP = Percentage of tweets containing terms in ggList
Using timeline retrieved earlier
Filter accounts with GP > threshold value
ggAccounts
For accounts in ggAccounts
Create user reference graph
Sort nodes by their in-degree in decreasing order
Sequentially contact each user and request them to be interviewed for the study
Users who agree to participate
UOB participants

Fig. 3. Steps taken to recruit UOB participants

outside the US, had fewer than 20 followers, fewer than 100 tweets, or more than 10,000 tweets (these are often bot accounts). We call the list of remaining Twitter accounts, *blockedAccounts*.

We retrieved the Twitter timelines of *blockedAccounts*, and constructed a corpus of words that combined tweets from these timelines. We used a list of stopwords to filter out terms like ‘the’, ‘at’, ‘on’ from this corpus [4]. Next, we created a list, *tfList*, by arranging words from this corpus in decreasing order of their term frequency. We manually inspected the first 500 words in *tfList*, and extracted a list of terms related to Gamergate. We called this extracted list *ggList*, and it contains terms like ‘#Gamergate’ and ‘#sjwtears’.

For each account in *blockedAccounts*, we calculated *GP*, the proportion of all tweets containing any of the terms in *ggList*. We filtered out the accounts having *GP* below a fixed threshold level, *t = 0.3*. We also filtered out accounts not posting in English. We called the list of remaining accounts, *ggAccounts*.

The user reference graph

Next, we built a directed graph of accounts in ggAccounts, and found the accounts most central to this network. We treated each account as a node in this graph. If a Twitter account *a* mentioned[9] another account *b* (using @[handle]) in any of his posts, we added a directed edge from *a* to *b* in the graph.

We collected a ranked list of 100 nodes having the highest in-degrees in this network. These nodes represented accounts that are expected to be heavily invested in the Gamergate movement, and influential among the blocked accounts on Twitter. We then sequentially contacted these users on Twitter to recruit them for interviews. Figure 3 describes this process. Note that although we recruited users associated with GamerGate as part of a broader research effort, we do not discuss our findings related to GamerGate in the current paper. We focus on our findings on online harassment and blocklists in this paper.

3.1.2 Selecting SB Participants. We used Twitter again to select interview participants who subscribed to GGAB. We collected all accounts that followed "@ggautoblocker," the official Twitter account of GGAB. Following this, we used a process similar to the one in the previous section. We filtered irrelevant accounts and retrieved the Twitter timelines of the remaining accounts. As before, we created a Twitter network of these accounts using their mentions, and curated a ranked list of users central to the network.

We then contacted the users on this list by messaging them on Twitter. Since individuals who follow the GGAB Twitter account don't necessarily subscribe to the GGAB blocklist, we asked all the potential interviewees whether they had subscribed to any blocklist. We only interviewed users who had subscribed to at least one blocklist.

3.2 Interviews

As discussed above, for each of the two groups, we invited the sampled users to participate in semi-structured interviews with us by contacting them on Twitter. About one in every five users we contacted agreed to do the interview with us. In all, we conducted 14 interviews with each of the two groups. Participation was voluntary, and no incentives were offered for participation.

The interviews began with general questions about which SNSs participants used, and their pros and cons. This provided necessary context to ask the participants about specific moderation problems and the use of Twitter blocklists. After developing some rapport with the participants, we asked them questions about their personal experiences and perceptions of online harassment. The interviews for the two groups followed different interview protocols. We conducted interviews over the phone, on Skype, and through chat, and each interview session lasted between 30-90 minutes. Some participants were contacted for brief follow-up interviews, for further clarification.

3.3 Participants

Most of the participants in our study reported being in their 20's and 30's. Among blocklist subscribers group, seven participants reported being male, four reported being female, two identified as transgender females and one participant identified as non-binary. We read online postings and

[9] A mention is a post on Twitter that contains another user's @username anywhere in the body of the post [48]. Responses to another user's tweet are also considered as mentions.

Table 1. Blocklist Subscriber Participants

ID	AGE	GENDER	CISGENDER/ TRANSGENDER	OCCUPATION	COUNTRY
SB-01	36	Male	Cisgender	Web developer	USA
SB-02	21	Female	Transgender	Student	USA
SB-03	49	Male	Cisgender	Software engineer	USA
SB-04	24	Female	Cisgender	Student	USA
SB-05	24	Male	Cisgender	Courier	USA
SB-06	23	Male	Cisgender	Student	USA
SB-07	36	Male	Cisgender	Academic	Australia
SB-08	22	Female	Cisgender	Student	USA
SB-09	31	Female	Cisgender	Student	UK
SB-10	42	Male	Cisgender	IT consultant	UK
SB-11	41	Female	Cisgender	Physics instructor	USA
SB-12	26	Female	Transgender	Call center employee	USA
SB-13	27	Other	Not available	Student	Canada
SB-14	23	Male	Cisgender	Unemployed	Germany

Table 2. Participants blocked by blocklists [a]

ID	AGE	GENDER	OCCUPATION	COUNTRY
UOB-01	38	Female	Childcare worker	USA
UOB-02	-	Male	Software developer	USA
UOB-03	28	Male	Consultant	USA
UOB-04	27	Male	PC repair	USA
UOB-05	36	Male	Medical professional	USA
UOB-06	33	Male	Game designer	Italy
UOB-07	20	Male	Student	UK
UOB-08	38	Male	Caregiver	UK
UOB-09	32	Male	Self-defense instructor	USA
UOB-10	32	Male	Appliance repairer	Mexico
UOB-11	33	Male	Game designer	USA
UOB-12	40	Female	Writer	UK
UOB-13	32	Male	Teacher	Netherlands
UOB-14	33	Male	PC repairer	USA

[a] All participants in this group are cisgender.

profile details of our participants on Twitter, and they indicated that some of the participants who identified as female in our interviews were also transgender. Among the participants who were put on blocklists, twelve identified as male and two as female.

Participants were self selected: we interviewed users who agreed to talk to us. Although most of our participants are from the US, we also had participants who live in Australia, UK, Canada, Germany, Italy, Mexico and Netherlands. The interviewees included a blocklist creator and a blocklist moderator. Tables 1 and 2 provide some demographic information about our participants.

3.4 Analysis

We transcribed data from our interviews and read it multiple times. Next, we conducted an inductive analysis of these transcripts. We summarized our data with open codes on a line-by-line basis [8]. We used the MAXQDA qualitative data analysis software (http://www.maxqda.com) to code our transcripts. Next, we conducted focused coding by identifying frequently occurring codes in our data and using them to form the higher-level descriptions. We then engaged in memo-writing and the continual comparison of codes and their associated data with one-another. We conducted iterative coding, interpretation, verification, and comparison through the course of the research. The comparisons led to the formation of axial codes that described seven overriding themes. In addition to the ones reported in the paper, themes such as different perspectives on GamerGate movement emerged but were excluded in further analysis. Finally, we established connections between our themes and these connections contributed to the descriptions of phenomena that we present in our findings [8].

3.5 Researcher Stance

The issues of online harassment and content moderation are sensitive, and as authors of this paper, we think it is important that we reflect on our position in this space. Our previous research on these subjects has shaped our perspectives on the current work. We identify online harassment as a systemic problem in the realm of the Internet, and like many other systemic social issues, it disproportionately affects women and other marginalized groups. We share the conviction that urgent efforts need to be made to spread awareness about the extent and severity of the consequences of online harassment. While we respect every individual's right to freedom of speech, we also recognize that abuse and harassment should not be justified in the name of free speech. We further believe in the need for designers and researchers to create tools that provide victims of online harassment the necessary support and security. However, our stances have developed through the course of this study and we have come to see that such technologies can also carry the risk of falsely accusing individuals of harassment.

Our goal in this paper has been to investigate the effectiveness of one anti-abuse tool in addressing online harassment in a fair way. We have not evaluated the public or private communications of our participants with other users. Therefore, we are not in a position to pass judgment on whether online harassment occurred or did not occur in different contexts. However, our methods have allowed us to listen to our participants on both sides - those who used this tool to avoid being harassed as well as those who were identified as harassers and blocked by the tool - and understand their views on these complex issues. Therefore, we present our findings as subjective perspectives of our participants. With our analysis, we hope to inform the readers about the complexities and challenges of online moderation.

4 FINDINGS : ONLINE HARASSMENT

4.1 Different perceptions of online harassment

In this section, we discuss perceptions of online harassment from two sides: users who have subscribed to blocklists (SB users) and users who have been blocked on GGAB blocklist (UOB users). As we discussed in Section 2.1, online harassment is not defined specifically and it can be difficult to distinguish harassers from non-harassers. By talking with both sides, a more nuanced narrative emerges than a simple contrast of good and bad actors.

Although the perceptions of online harassment generally vary with the users' overall experiences, many of our SB participants mentioned being disturbed, and in some cases, traumatized, by online abuse. Participant SB-11 said that she had to start taking anti-depressants in order to cope with harassment. Describing an incident in which she found doctored photos of herself, she said:

> *"They were extreme. Extreme. Violent, and things that just stuck in my head that I couldn't ... they weren't just gross. They were violent. I couldn't shake them. I had to take a break and they kept intrusively coming into my thoughts. It was really awful." – SB-11*

Our participants characterized online harassment as acts ranging from someone posting a spoiler about the new Star Wars movie to someone sending them death threats. Four of our participants mentioned that someone had tried to get them fired from their job by contacting their place of work because of online disagreements.

Many users of the UOB group did not realize that online harassment can have serious consequences. A few UOB users said that they don't believe that online harassment is a legitimate problem because they can block or mute anyone that bothers them. Some participants proposed that online harassment shouldn't be taken too seriously. Participant UOB-06 said:

> *"It's certainly unpleasant but it has nothing to do with terms like "oppression" and "danger" that often get thrown around. I am an LGBT rights activist in Italy and I have met people that face some real oppression and danger in their lives. . . I find that every form of oppression that can be filtered out or avoided by closing a browser is more like an annoyance than a problem." – UOB-06*

Some participants told us that different users have different sensibilities, and often, individuals view the same discussion in very different ways because of the differences in their points of view, identities, or the issues that they are tuned into. They argued that these differences may contribute to some users perceiving that they are being harassed in situations that others consider as an expression of valid political speech. Furthermore, a few UOB participants noted that they have seen instances in which disagreements on an issue are deliberately portrayed as harassment by others. They considered such cases a strategy by their opponents to push a political agenda.

> *"There's a narrative that social justice ideologues try to push. That, for example, if you think the female pay gap is a myth (or even not as severe as [what] third-wave feminists say), [they claim] that you're sexist, misogynist, and if you're a woman, you have internalized misogyny." – UOB-11*

Some SB participants said that there are instances when the person doing the harassment doesn't fully realize the extent of impact of their acts upon the harassed users. Other SB users felt that harassers are often gullible individuals who are dis-informed about political issues. For example, Participant SB-01 believes that some GamerGate supporters become aggressive in their responses because of their basic misunderstandings about the nature of journalism.

> *"GamerGate supporters ally with hatemongers because they too feel like outsiders, like they're ignored, and joining a mob is the only way to get the narrative centered around them...Their anger is genuine, even if the narrative is false." – SB-01*

In contrast to the perspectives in the previous section, UOB participants felt that many online commenters overreact to trivial cases of perceived offenses. They also worried about the dangers of backlash against such overreactions. Participant UOB-02 felt that some users who promote socially progressive views and stand against online harassment actually hurt their own cause by dismissing anyone who disagrees with them on any issue:

> *"There are these kinds of social justice issues out there that are really, really important and that really address a lot of very real marginalization and just horrible things that are going on and I feel like to some extent, some of these bad actors in that space have*

> *kind of sullied the name of something that should be a lot more compassionate than it currently is." – UOB-02*

Some UOB participants felt that the media often highlights the online harassment of a few groups while ignoring similar abusive behavior against opposing groups. Participant UOB-04 expressed sympathy for the targets of harassment but felt that it was duplicitous of media and many people with authority to depict different instances of online harassment in ways that he considered biased:

> *"It is hypocritical to me, though, because while I often see one specific group of people and the issues they face being portrayed as harassment (and this group of people are often friends of the people doing the portrayal) - I see other groups of people who face similar hardships being written off and undermined by the same group." – UOB-04*

Participants noted that prominent perpetrators of harassment include groups ranging from GamerGate supporters and GamerGate opponents to radical feminist groups. Some participants pointed out that harassers also include trolls who conduct harassment as a kind of cultural performance art. This is similar to what Whitney Phillips found in her work on trolls in which she provides an empirical account of the identities, attitudes and practices of trolls, and their impact on the digital media environments [37]. She argues that "the vast majority of trolling is explicitly dissociative…the mask of trolling safeguards trolls' personal attachments, thereby allowing the troll to focus solely on the extraction of lulz[10]" [37].

A few of our SB participants believe that, in contrast to trolls, there are harassers who develop an emotional investment in hurting their targets. For example, Participant SB-11 told us that harassers include users with serious psychological challenges who deal with their personal traumas by attacking others online. She argued that such harassers exhibit characteristics that are quite distinct from trolls, for example, they often don't have their identity anonymized on social media. She distinguished such users from trolls by saying:

> *"Some of the stuff they (trolls) said was kind of shock value…they weren't obsessed with me. They didn't have some sort of emotional investment in hurting me. When you find those people who are really invested in you personally, for whatever reason, God knows why, that's the scariest ass thing." – SB-11*

Some UOB participants complained that they were perceived as harassers by other users because of their mild association with controversial individuals on Twitter, and not because of their own activities. Participant UOB-01 told us that she was harassed by many users and was accused of being a "gender traitor" after she was put on GGAB blocklist. She questioned the decision-making process of GGAB moderators and felt that they did not have any qualms about blocking the users who don't harass:

> *"I've always tried to talk to them but they simply don't care if you're a good person. If you don't agree with their ideals, you're automatically the bad guy…I tried to get removed [from a blocklist] and was denied because I retweeted people like Totalbiscuit and Adam Baldwin. I was guilty by association." – UOB-01*

4.2 Tactics used by harassers

In this section, we discuss some behavior patterns and tactics that our participants identified as manifestations of online harassment. Table 3 lists these tactics among others and briefly defines them.

[10] "lulz" is a corruption of lol (laugh out loud) that signifies unsympathetic laughter, especially one that is derived at someone else's expense [37].

Table 3. Tactics used by harassers

Tactic	Description
Brigading	A large number of users, often those belonging to the same group, posting together on other online spaces in order to disrupt conversations.
Concern trolling	Visiting a site of an opposing ideology, and disrupting conversations or offering misleading advise in the guise of supporting that ideology.
Dogpiling	Many users posting messages addressed to a single individual. The intent of any sender may not be to perpetrate harassment, but it results in the targeted individual feeling vulnerable.
Dogwhistling	Using messages that sound innocuous to the general population, but have special meanings for certain groups. Such messages are used as a covert call to arms to target an individual or a group.
Doxing	Revealing someone's private information online with an intent to intimidate them or make them vulnerable to offline attacks.
Identity deception	Providing a false impression of one's own gender, race, etc. to gain advantage in online conversations.
Multiple SNSs	Using multiple social network sites to retrieve more information about the targets.
Sealioning	Politely but persistently trying to engage the target in a conversation. Such conversations are often characterized by asking the targets for evidence of their statements.
Sockpuppeting	Using an alternate account to post anonymously on social media. This is often done to feign a wider support of one's own postings.
Subtle threats	Using subtle hints to intimidate targets and make them aware that their personal information is exploitable.
Swarming	A group of users simultaneously attacking the same individual.
Swatting	Anonymously contacting and misleading law enforcement to arrive at the unsuspecting target's address.

4.2.1 *Subtle threats.*

Some participants argued that often, the perception is that online harassment is transparently malicious, involves violent threats, etc. but online harassment can manifest more subtly too. For example, Participant SB-01 said that he received messages from strangers, which indicated that they had gleaned a lot of personal information about him from his social media postings.

> *"Some strategies I have noticed: [messages like] 'Paul [11], you work in tech in Portland, how can you be so ignorant about this issue?' Another common example was mentioning to me that I have kids or making comments on selfies I'd shared a few weeks ago." - SB-01*

Participants believe that such messages intend to make the recipients aware that the harassers have read through many of their posts, and that their personal information is exploitable.

4.2.2 *Using multiple social networks.*

A few participants reported that some trolls had gone through their profiles on multiple social media sites in order to gain personal information about them.

[11] Name changed to preserve anonymity

> *"A few days ago, I had a men's right activist who I blocked on Twitter and then he went and found my Facebook account and sent me threats through Facebook." – SB-07*

Some participants also noted that a few troll groups organize harassment on other, more obscure websites, but carry it out on more popular social media platforms like Twitter, taking advantage of their ineffective moderation.

4.2.3 Dogpiling. Some participants said that sometimes, the harassment is because of the volume of tweets. They referred to such cases as 'dogpiling'. They felt that in such cases, the intent of many posters may be that they want to debate someone who they don't know, but whose post they come across. Participant SB-11 said that in such cases, "when you're getting like a hundred messages all wanting to debate you, then it feels like you're being overwhelmed." Other participants felt that such dogpiling occurs as a result of coordinated troll campaigns.

> *"The first mass contact would be like shining a spotlight on you, and then all that other stuff would happen - digging in, contacting your family, or your employers" – SB-02*

Some participants noted that dogpiling occurs on Twitter when one of the participants in a discussion has a large number of followers who interject themselves in the conversation.

> *"Somebody was bothering J.K. Rowling and saying rude things to her, and she responded with something like, "Yeah, but your screen name is stupid," or whatever, right? Because J.K. Rowling has millions of followers, this person just got descended on." – SB-02*

Participants said that in such incidents, the individual with many followers may or may not have the intention to dogpile a target. When such dogpiling is deliberate, this phenomenon is referred to as "dogwhistling" (see Table 3).

4.2.4 Identity Deception. Some previous studies have noted that trolls and harassers engage in identity deception [14], and its varieties such as gender deception [46] and age deception. Our participants also consider such behaviors as manifestations of online harassment.

A few participants told us that in some cases, harassers use identity deception to strengthen their argument in discussions. They present themselves as belonging to a minority group or an oppressed community that they don't actually belong to. Participants felt that this is done so that users who wish to be sensitive to the opinions experienced by minority groups don't contradict them.

> *"Those that come out and say I'm a 13-year-old trans-girl from this Christian conservative family and I'm having such a hard time, and you know what I think? Then they just go on some rant, strong argument rant about something to try to prove that all these social justice people are completely ridiculous and a lot of people go - I'm not going to call out a 13 year old girl because that doesn't feel right." - SB-11*

Participants noted that trolls also frequently use identity deception to harass others. For instance, some GamerGate supporters claimed that a few troll groups incited both GamerGate supporters as well as opponents by posing to be on the other side, and posting offensive messages.

Some participants said that many harassers and trolls create multiple accounts, and use them to overwhelm their targets. In some cases, multiple accounts are used to pretend that other users agree with and support their abusive responses to the target. Participant SB-10 said that when an account gets banned on Twitter, the abusers often quickly make a new 'sock' account, and resume attacking their targets (See "sockpuppeting," Table 3).

4.2.5 Brigading . Brigading refers to a concerted attack by one online group on another group, often using mass-commenting or down-voting [24] . Some of our participants who actively use Reddit noted that on Reddit, brigading frequently occurs on subreddits with opposing ideologies. When someone posts on subreddit r_1 a link to a submission or comment c posted on a different subreddit r_2 with the knowledge that c would be unpopular on r_1, it has the effect of r_1 users down-voting c on r_2 and posting replies to c that are contrary to the values of r_2 users.

Some participants noted that on Twitter, brigading often occurs through malicious misappropriation of hashtags. They mentioned that trolls often "brigade" hashtags that are used by minorities, and disrupt their conversations.

> *"Many of them would flood tags like #ICantBreathe in support of Eric Garner and other victims of police violence with racism and gruesome images or pornography." - SB- 13*

Some users noted that brigading is sometimes used to spread misinformation. They also pointed out that feminist hashtags like #YesAllWomen, tags used by disability advocates, and tags that abuse victims use to share their stories are also frequently brigaded.

4.2.6 Sealioning. A few participants told us that harassers, particularly those who support GamerGate, engage in persistently but politely requesting evidence in a conversation, a behavior referred to as "sealioning." The name "sealioning" is derived from a comic in which a sea lion repeatedly tries to talk to a couple and annoys them [30]. Sealioning is viewed by many users as intrusive attempts at engaging someone in a debate.

4.2.7 Doxing. Doxing refers to revealing someone's private information online. Our participants told us that doxing often occurs on relatively obscure internet forums that are dedicated to doxing, but it also occurs on more popular social media platforms. A number of participants from both groups mentioned that they were doxed online or had witnessed doxing. Doxing a user using a pseudonym can be damaging to that user, because it may reveal information about that user that he may not be comfortable associating with publicly. Often, doxing involves revealing information such as social security number or residential address. In such cases, many fear that the doxed person's personal safety can be endangered.

Some participants expressed frustration that their reporting of doxing is met with the platforms' response that the doxed information is available through a search of their username on internet search engines, and therefore, it cannot be removed because it is public information. Others complained that the platforms respond to doxing only if the person being doxed is a public figure.

Other privacy intrusions identified by our participants include attempts to hack online accounts and calling employers to try to get the targets fired. They also noted that in many cases, threats of these actions are used to harass individuals.

4.3 Who is vulnerable to harassment?

4.3.1 Perceiving minority groups to be more vulnerable. Online harassment knows no boundaries, and virtually anyone can become a target. However, some of our SB participants insist that certain groups like the transgender community and Muslims are especially vulnerable to online harassment. Some participants noted that females, particularly women of color and feminists, are another vulnerable group. Many participants felt that the identity of harassed users plays an important role in how much and in what ways the harassment is conducted. Others pointed out that they enjoyed privileges due to their gender, race or nationality, and were not very vulnerable to harassment.

4.3.2 *Believing that dependence on online communities for social support increases vulnerability.*

Our transgender participants told us that the online transgender community on Twitter is very important to them, because it allows them to connect to individuals with similar experiences.

> *"I'm transgender, something it's taken me years to come to terms with. And then over the past couple of years, I've learned that there's a whole community of people a lot like me. There's a lot of bonding." – SB-12*

They felt that the dependence of transgender users on their online community as a primary means of social support makes them more vulnerable, because it is more difficult for them to leave the platform. Participant SB-12 mentioned that some of the harassment of transgender users comes from other users within their community.

> *"Trans people, especially trans women, are often harassed online. As much as it pains me to say it, we aren't always good to each other, either. I've seen trans women make death threats toward each other." – SB-12*

4.3.3 *Believing that a decrease in anonymity increases vulnerability.*

Early research on online communities has shown that the relatively open nature of the information on SNSs and many users' lack of concern with privacy issues expose users to various physical (e.g., stalking) and cyber (e.g., identity theft) risks [21]. Users who have taken great efforts to create an online presence on a platform find it difficult to leave the platform [14] if they face abuse, and are therefore vulnerable. Some of our interviewees also said that there is a connection between anonymity of a user and his or her vulnerability to harassment.

> *"I found that the less anonymous you are, the more cruel people can be." – UOB-01*

Some participants felt that revealing too much information about oneself or being too outspoken about sensitive issues on some platforms can be dangerous. However, they also found it difficult to determine exactly how much personal information is "too much" information in a given context.

4.4 Support of harassed users

In their study of Hollaback [12], a social movement organization whose mission is to end street harassment, Dimond et al. found that sharing and reading others' stories of how they define harassment and respond to it helped shift harassed individuals' "cognitive and emotional orientation towards their experiences" [13]. We found evidence of participants drawing comfort from sharing their experiences in our study too. Our participants appreciated the support that they received from other users on the platform during episodes of online abuse. For example, Participant UOB-01 described how she drew comfort from sympathetic messages from GamerGate supporters:

> *"Yes, I had random strangers tweet at me their support…Even when I had angered a few people by leaving GamerGate the way I did, I had many still in the movement show their support. Twitter has never been short of amazing people. It's just the louder, angrier voices carry more weight." - UOB-01*

Participant SB-11 discussed how many harassed users drew comfort from her volunteer work of blocking and reporting accounts who harass them online.

5 FINDINGS : BLOCKLISTS

We discuss our findings on blocklists in this section. Our discussion of different perceptions of online harassment in Section 4 provides context for understanding different views of the use of blocklists in this section.

[12] https://www.ihollaback.org

5.1 Algorithmically curated versus socially curated blocklists

Some popular blocklists (e.g., GGAB) are algorithmically curated while others (e.g., Block Bot) are socially curated by a small group of volunteer users. When talking about the different blocklists that are popular on Twitter, Participant SB-11, a Block Bot moderator, drew distinctions between algorithmically and socially curated blocklists. She said that although algorithmically curated blocklists cannot make the complex decisions that humans can make via a socially curated blocklist, their decisions may be perceived by blocklist users as more objective, as they have predefined, fixed criteria. On the other hand, Participant SB-11 further argued, socially curated blocklists would tend to have fewer false positives [13] because they are curated by humans.

Block Bot, in particular, blocks a Twitter account when two different moderators agree that the account deserves to be blocked. Additionally, every Block Bot moderator has the right to unblock any blocked account. Block Bot moderators believe that this reduces the probability of having false positives. Different users may have different views of who should and who should not be blocked, and therefore, it is difficult to predict how many Block Bot subscribers would agree with its moderators' decisions. We note however that many participants complained about the high number of false positives on GGAB, an algorithmically curated blocklist, but we didn't hear such complaints about Block Bot in our interviews.

> *"When Harper[14]'s GGAB was made, it was so broad that it had company accounts on there, e.g., KFC twitter [account], which follows back people who follow it."* – SB-07

5.2 Why do users subscribe to/avoid subscribing to anti-harassment blocklists?

When we asked SB participants how they came to start using blocklists, some users in the sample mentioned that they subscribed to blocklists after receiving targeted harassment. They pointed out that often, these harassers belong to the same group such as GamerGate or TERFs (trans-exclusionary radical feminists). Some users noted that they started using blocklists because reporting abuse to Twitter proved to be futile.

> *"It would have been early last year when I sent a tweet using the #GamerGate, and I compared GamerGate to men's rights activism...I received hundreds of tweets every hour, and I discovered that there were discussion boards where I've become the focus of discussion among people participating in GamerGate...That went on for about a week or so until I was able to install an auto blocker and basically just cut them out, so they couldn't continue the harassment."* – SB-07

> *"I eventually just grew tired of trying to either respond to these people and yeah, eventually, I just didn't want to talk to these people anymore, because it was a repeating pattern...They say the same offensive remarks, same insults, same aggressive behavior. It wasn't really changing anything when reporting them to Twitter. "* – SB-14

Other participants preemptively subscribed to the blocklists because they thought it would shelter them from seeing hate filled content. Some participants assumed that users who are on blocklists aimed at blocking a specific deviant behavior would be bigoted in other ways too.

[13] Throughout this article, we use 'true positives' and 'false positives' in the context of blocklists metaphorically in order to refer to accounts that most subscribers of the given blocklist would prefer to block and not block respectively. We do not make any claims about whether the 'true positives' are genuine *harassers*. Making such claims would require operationalizing what exactly counts as harassment in the analysis at hand, which is beyond the scope of this paper.

[14] Randi Harper is a Software Engineer, based in Portland, Oregan. She was involved in the GamerGate controversy and has created a few open-source anti-harassment tools including GGAB [36] .

> *"A lot of people are actually on multiple of these block lists because a lot of the anti feminist and the GamerGate people are one and the same. Also a lot of racists are also one and the same people. It's mostly a very similar kind of people." – SB-14*

Many users who subscribed to blocklists also promoted the use of blocklists on Twitter and other social media as a solution to mass harassment on Twitter. Participant SB-05 said that he followed many individuals who were harassed by the GamerGate movement, and who used and promoted GGAB. This convinced him to preemptively start using GGAB in 2014. Some users told us that the use of anti-abuse blocklists was so popular among pro social-justice users, that anyone who followed such users knew about blocklists.

> *"Nobody really questions the necessity of GGAB, since [Gamer]gators in your mentions are like an STD. You did something wrong, and you need to get rid of them." – SB-12*

Some participants took to using blocklists after they failed to find a common ground with groups of users with opposing views. Participant SB-12 mentioned that she used GGAB because after having a few discussions with GamerGate supporters, she realized that neither she nor the GamerGate supporters had anything to gain from listening to one another. Similarly, Participant SB-01 said:

> *"I was getting constant mentions from GamerGate accounts, and I was wasting tons of energy replying. When someone asked "what have you been up to?", all I could think of was: arguments with anonymous neofascists. I saw a few people discussing preemptive blocking and looked into it." – SB-01*

In addition to GGAB and Block Bot, some participants also use other anti-abuse blocklists. For example, Participant SB-08 uses a blocklist compiled by her and her friend in addition to GGAB. Participant SB-01 uses a blocklist manually curated by a widely-followed artist. Participant SB-14 maintains his own private blocklist and shares it with a few of his friends.

None of the UOB participants subscribe to any anti-abuse blocklists like GGAB and Block Bot, often citing the reason that they are opposed to avoiding discussions with those they disagree with. Some SB participants also deliberately avoided using blocklists despite claiming that they suffered online harassment. For example, Participant SB-03 described why he doesn't use any blocklists:

> *"I find it useful to follow some of the ringleaders of these brigades to keep track of them...A disadvantage of using blocklists is that you may miss early warnings, or actual engagement." – SB-03*

Participant SB-13 used three blocklists but he avoided using GGAB blocklist because he felt that he wasn't high profile enough to be a direct target of GamerGate supporters, and he could avoid becoming a target just by not using the hashtag #GamerGate.

5.3 How did user experience change after using anti-abuse blocklists?

Many SB users said that their Twitter experience improved after they started using anti-abuse blocklists. They told stories about how they stopped getting large numbers of unwanted notifications, and tended to get only genuine inquiries from strangers instead of abusive messages.

> *"It does quite a good job of putting up a firewall between me and sections of Twitter that I don't really want to have access to my content and I don't really want to engage in dialogue. By and large, I've found that it's a pretty effective way of taming Twitter." – SB-07*

> *"Twitter is a cleaner place when I go looking at things. I don't see the thousands of things from racist, bigoted people. It's just nicer for me to not have to see terrible stuff written on a daily basis, which is good." – SB-08*

Participant SB-10 felt that a few minority groups, for example, transgender communities, especially benefit by the use of anti-abuse blocklists:

> *"I certainly had a lot of transgender people say they wouldn't be on the platform if they didn't have The Block Bot blocking groups like trans-exclusionary radical feminists – TERFs." – SB-10*

Participant SB-02 said that she noticed a decrease in unwanted notifications in stages. When she first subscribed to a blocklist, her notifications decreased a lot and then leveled off. After some time, such notifications decreased further. We suspect that this decrease in notifications could have corresponded to the periodic blocking of new accounts by the blocklist.

Many participants found the use of blocklists liberating. Participant SB-11 said that the use of blocklists empowers individuals to set their own boundaries by providing them with effective tools to control the content they consume. None of the participants who used blocklists seemed to be bothered by whether blocklists blocked users who they would not have blocked. For example, Participant SB-09 said that she noticed a few accounts that were blocked because of the blocklist's false positives and she manually unblocked these accounts. However, the benefits of having fewer unwanted notifications and abusive messages far outweighed this cost for her. Many other participants shared a similar view:

> *"If I* really *want to know what blocked people are saying to me, I can go check from an incognito browser, but realistically, there are millions of people on Twitter. Saying I don't want messages from less than 1% of them isn't going to damage the range of people I can hear from." – SB-01*
>
> *"It's not 100% foolproof, but it's very effective. Sometimes I find people that I like that happen to have blocked, so I unblock them. Sometimes it's a little tedious, but it's worth it." – SB-06*
>
> *"It's not a perfect solution, but no anti-harassment tool is going to be perfect. I'm much more worried about the impact of targeted harassment on me and my career than I am about the minor inconvenience that someone might face because they are unable to contact me on Twitter." – SB-07*

Many participants who found blocklists useful still pointed out that they would much rather have Twitter address the problem of harassment effectively so they don't have to use third-party tools like blocklists. Not all the participants consistently found blocklists dependable. Although Participant SB-09 found blocklists useful when she first subscribed, she feels differently about them now:

> *"I'm far from keen on them, now that there are better tools out there, and after someone with a large following used them to cut a lot of innocent people off from a number of communities on Twitter by sharing her personal blocklist and claiming it was mostly 'MRAs [Men's rights activists] and trolls'" – SB-09*

Participant SB-04 felt that although using blocklists improved her Twitter experience slightly, she still suffered harassment.

> *"I still got harassed but my experience was probably slightly better because of it. It's not like it kept them all from making new accounts." – SB-04*

5.4 Challenges of social curation

5.4.1 Motivating moderators. Social curation is not a trivial activity. It requires many volunteer users contributing several hours each week and coordinating with one another to moderate sexist, racist and homophobic content. Reviewing rape and death threats, violent images, and aggressive threats over a long period can be psychologically damaging. Some media reports have described how regulating the internet can deeply affect moderators and even drive them to therapy [40, 52].

What then motivates the moderators of socially curated blocklists to continue blocking trolls and harassers for free? We asked a Block Bot moderator what drives her to continue moderating, and she replied:

> *"One thing that motivates me is that I know that we're providing service that people need to stay connected to others. Especially with the trans-community where maybe in your city, in your community, there's a handful of other trans-people. Otherwise, you aren't connected to people with similar experiences as you...a lot of people said [to me] point blank: If I didn't have your service, I couldn't be online...Now, I can, and so I keep [more] connected to the world more than I would be able to otherwise." – SB-11*

5.4.2 Guarding against rogue moderators. Beyond the problem of keeping volunteers motivated to continue moderating, socially curated blocklists have a number of challenges. Participant SB-11, a moderator for Block Bot, mentioned that in rare instances, one of the moderators of Block Bot acted irrationally, and blocked a number of people who didn't deserve to be blocked. Following this, the Block Bot team did some damage control, and put in technical fail safes. However, such incidents highlight the vulnerabilities that any socially curated blocking mechanisms can have.

5.4.3 Making decisions about perpetrators from vulnerable groups. Moderation decisions can be challenging for the moderators. When an account in question belongs to an abusive user from a vulnerable group, the decision of whether or not to block that account becomes difficult. For example, Participant SB-10 said:

> *"You might get an abusive member of the transgender community, and then the question is, do you block them and isolate them from their own community, given that the suicide rate in transgender people is actually very high. " – SB-10*

5.4.4 Moderators and users having different perspectives. Another challenge is that the moderation decisions of a socially curated blocklist are biased by the particular perspectives of the blockers. Everyone has different viewpoints and tolerance level, and what might offend one person may be perfectly reasonable to another. When users subscribe to the blocklist, they block accounts that are moderated through complicated decisions taken by the blocklist curators. This can be problematic because the users and the blocklist moderators may have very different definitions of what harassment entails. As Participant UOB-09 described, "You are guilty if they (blocklist moderators) say you are." Participant SB-11 explained:

> *"Most of the trans-people we've had as blockers are trans-women. They're going to have a certain perspective on what would be considered a blockable offense that's going to be slightly different than somebody else." – SB-11*

Other participants also worried about this problem, and felt that socially curated blocklists may be efficient only when they are curated to serve specific social groups constituting of individuals with similar ideas of what harassment means to them.

5.4.5 Resolving conflicts among moderators. Participant SB-11 told us that there are even differences among the Block Bot moderators on who they consider should be blocked. This can result in conflicts among the moderators.

> *"If somebody decides that this person shouldn't be on the [Block] bot, they just aren't. The only time that things have gotten really, really difficult is if there's somebody really close to our social circles who somebody feels strongly that their account should*

> *be placed on the Block Bot. That's happened a couple times and it was pretty terrible. We've avoided that sort of thing recently because you learn from when it happens." – SB-11*

5.4.6 Making trade-offs between being transparent and resisting attacks. The Block Bot moderators have made the list of blocked users publicly available on their website, www.theblockbot.com. They felt that having this list public has made the Block Bot more transparent as the potential subscribers can find out what they are signing up for by browsing the profiles of users who have been blocked. However, it has also made the moderators more vulnerable to attacks by users who are put on the blocklist. They said that they would rather ignore such users than annoy them or engage with them.

The Block Bot moderators save the tweets that led to each user being blocked and present them when any blocked user inquires them about the reason why he or she was blocked. They said that showing the posts that resulted in their blocking often results in convincing the blocked users to drop their appeal to be unblocked. Users who are put on the Block Bot are not informed that they have been blocked. Participant SB-11 described the implications of this decision:

> *"The pro was, of course, we're not interacting with them, we're not escalating anything. They might not even know the Block Bot exists and they've been on it for years. On the other hand, if we did make a mistake, somebody got on the Block Bot and they don't want to be on the Block Bot and they can't really talk to us about it if they don't know they're actually on it." – SB-11*

5.5 Perception that blocklists block too much/block unfairly

All the UOB participants felt that the currently popular blocklists block unfairly because they were surprised by finding themselves on blocklists and they did not feel that any of their actions warranted being blocked.

> *"If you suddenly get put on a list of "the worst harassers on Twitter" when you haven't said anything on the platform for years then you sort of want to know why." - UOB-08*

Some participants felt that the criteria for curation of some blocklists - such as blocking all accounts who follow specific Twitter handles - was too crude to be considered reasonable.

Many participants worried about the personal biases of the blockers who compile socially curated blocklists or the developers who design or code algorithmically curated blocklists. A few SB participants also shared similar concerns and chose not to use some blocklists because they disagreed with the politics of individuals who managed those lists.

> *"GG block list is run by someone who had some skewed ideas of a few things that I happen to disagree with. Pretty bad stuff." – SB-06*

Some UOB users said that they couldn't access the pages of popular public figures, artists, etc. because they were using the blocklists the users were blocked on. A few UOB users claimed that they suffered professionally because of being put on blocklists.

> *"At a certain point, the International Game Developers Association sponsored the [gg]autoblock[er] claiming it was a way to block the worst harassers of Twitter. Beside not being a harasser, I am a game designer so an association that is supposed to protect me was accusing me instead." - UOB- 06*

Participant UOB-10 said that many new blocklists copy the accounts already put on popular blocklists. This leads to many users being blocked by accounts who wouldn't block them otherwise. A single individual's personal dislike and subsequent blocking of a user can snowball and end up excluding that user from many important groups. We also found evidence of this phenomenon outside our interviews. For example, one blogger complained that an influential Twitter user put many transgender users on her personal blocklist over minor disagreements: *"When someone in a position of trust and power blocks many marginalized people over minor disagreement, then it disseminates distrust and removes avenues of communication for those marginalized people...All those people are now blocked by all the tech contacts, feminists, celebs that sign up to her list, as well as anyone signed up to their block list"* [42].

5.6 Feelings about being put on blocklists

When we asked UOB participants how they felt about being put on blocklists, their reactions ran the gamut from indifference and mild annoyance to disgust. Some participants did not feel very strongly about being put on blocklists. Others felt a little irritated that some users would consider them "horrible" without asking for any proof just because they are put on the blocklists. Still others felt okay with it because they believed that any Twitter user had the right to block whoever they want.

> *"We just sort of laughed at it and shook our heads since it seemed like a dumb thing to waste time on." - UOB-14*

Participant UOB-01 said that she was put on a blocklist just because of being part of GamerGate, and her efforts to be put off the list were futile.

> *"I never tweeted anything mean at any one. I always said harassment was wrong. It got me nowhere." - UOB-01*

Many UOB users described similar incidents of being put on a blocklist unfairly. They felt that they were victimized just for having wrong associations on Twitter, and that they did not deserve to be perceived as harassers.

A Block Bot moderator told us that some users got really upset about being put on Block Bot, because they thought that the blocklist was making an incorrect claim about the kind of person they are. Some UOB users criticized people who subscribed to blocklists. They characterized the blocklist subscribers as individuals who are not open to challenging their own conceptions or questioning themselves.

> *"If someone needs to insulate themselves via lists someone else created, they have bigger problems than getting offended by what I have to say." - UOB-12*

A few participants felt that it is unfair to be blocked for disagreement on just one topic of discussion, for example, their support of GamerGate. Participant UOB-06 said that it is cowardly of blockers to block the accused who then have no means of responding to the blockers' allegations.

5.7 Appeals procedure

Some SB participants said that they did not feel very strongly about the necessity of a fair appeals process.

> *"Being blocked on Twitter is not a legal issue, it's not a censorship issue, it's not a human rights issue...If there is a way for people to appeal, that's great. I don't actually think it's compulsory, to be honest." - SB-07*

Many UOB participants expressed an indifference about using an appeals process. For example, Participant UOB-12 said that it never occurred to him to try to get himself off the blocklists.

> *"People aren't obliged to speak to me. It's still a (fairly) free internet." - UOB-12*

Some UOB users were not aware of the existence of an appeals procedure that they could use to get off the blocklist. A few participants mentioned that they were discouraged from appealing to get off the blocklists because they had seen discussions about many cases of such requests by other blocked users proving futile. Participants UOB-01 and UOB-13 said that they sent messages to get themselves off the blocklists, but their requests were denied or they never received a response. This further contributed to their negative views of the use of blocklists.

> *"If you look at the behavior of the people who control these things, I think you'd have to be incredibly optimistic to expect a response, unless you are someone famous or rich or whatever." - UOB-13*

Many users said that they considered appealing to get off a blocklist to be too much of a bother and not worth the effort required.

> *"I don't really need an appeals process; again, that would be too much work. It's only if everybody I started following would block me because they were using the same block list, then I would go to it because I'm like, I'm not that bad." - SB-08*

This suggests that the existence of a fair appeals procedure may become more critical to those who are blocked if the number of users subscribing to a blocklist increases or if a blocking contagion occurs.

6 DISCUSSION

Social media allows users having different experiences, ideologies, and political opinions to interact with one another. Twitter, in particular, as a result of its open design, allows users to find conversations on diverse topics, and respond to any posts directly. This provides an extraordinary opportunity for constructive discussions, understanding different perspectives, and discovering bipartisan solutions to complex societal problems. Unfortunately, in many instances, the interaction of users with opposing viewpoints results in aggressive behavior. In this section, we use the findings from this study and expand on previous work to propose tools and interventions that designers and policy makers should consider in order to help cultivate civil discussions, and reduce instances of and mitigate harm from online harassment.

6.1 Focusing on vulnerable groups

Some users don't realize that their experiences may be quite different from other individuals. Our findings in this study show that many users don't grasp the emotional toll that their facetious remarks can exert on other users (Section 4.1). As Whitney Phillips describes, "even the most ephemeral antagonistic behaviors can be devastating to the target, and can linger in a person's mind long after the computer is powered down. This is especially true if the target is a member of a marginalized or otherwise underrepresented population, whose previous experience(s) of abuse or prejudice may trigger strong negative emotions when confronted with nasty online commentary" [37].

Our findings also suggest that the identity of harassed users plays a role in making them vulnerable to online harassment (Section 4.3). This provides the following opportunities for researchers:

6.1.1 Study oppressed groups and talk to them. Efforts to understand the specific needs of each oppressed group on an SNS, for example, through surveys or interviews of individuals from the group, can inform the modification of existing moderation mechanisms so as to better serve that group. Researchers can study the online activities of specific oppressed groups, and characterize the posting activities that lead to unusually high abusive responses. These findings can then be used to inform the group of the type of postings that have inadvertently invited abusive responses in the past. This knowledge may help such individuals make informed decisions on whether to engage themselves on certain topics.

In an ideal world, everyone would always be able to speak their mind without fear of harassment. It's important for system designers to work to achieve that goal. It would be unfortunate if marginalized groups learned to self censor. In the real world, however, consequences of certain kinds of speech can have a negative impact on individuals. Until the fundamental conditions that make the internet so conducive to harassment are ameliorated to some degree, it is strategic for individuals to at least be aware that something they are about to post has a high likelihood of provoking harassment. If we could create tools to alert individuals to that possibility, they could make a more informed choice about whether the benefits of particular speech outweigh the risks. Such a tool ideally could give people guidance on how to express the same ideas in ways that are less likely to attract abuse, or better still in ways that are more likely to be truly heard by the intended audience. We imagine such a tool could be of great use to individuals on both sides in our study.

6.1.2 Develop tools and systems that serve the special needs of vulnerable groups. Language and actions that are abusive to a particular vulnerable group, for example, transgender users, may not be offensive to other users. Therefore, researchers and designers may need to develop specialized tools and systems for each group. User studies employing the individuals from the target group as participants who use these systems can be deployed to evaluate the effectiveness of such systems.

6.2 Designing support systems for harassed users

Platforms can also design to support users who suffer online harassment. Aggressive behaviors can be diffused by providing users with an alarm functionality that alerts their friends of their need for help. Support systems can also help users who have suffered similar abuse connect to one another, and share their experiences. Our findings indicate that harassed users value messages of support during episodes of online abuse, even from strangers (Section 4.4). Therefore, systems that allow targets of harassment to receive support messages from other harassed users can help them cope with online harassment. In their study of Heartmob[15], Blackwell et al. argued that public demonstrations of support not only provide validation for targets of harassment, but also create powerful descriptive norms that help other users determine what behaviors are and are not appropriate in an online community [5]. We note that although support systems can provide critical support to harassed users, they are vulnerable to attacks by trolls and they need to be carefully designed to ensure that they are not misused.

Some harassed users may not be aware of how to use the tools on SNSs available to them. Platforms should provide guidelines and tutorials to their users on how to safeguard their privacy and use the anti-abuse tools available on the site. SNSs should also promote online resources like HeartMob so that the targets of online abuse can get information on supportive organizations and other helpful resources. Such measures would indicate to the harassed users that the platform is committed to addressing abusive behavior, and encourage them to not leave the site.

6.3 Improving blocking mechanisms

There are a number of ways that blocking mechanisms can be redesigned so that they better serve the needs of different user groups. Our findings suggest that there is a need for decentralized blocking mechanisms like Twitter blocklists that operate separately from the centralized moderation provided by Twitter. However, certain measures need to be taken to ensure that these lists block fairly as well as serve their subscribers appropriately.

[15] https://iheartmob.org

6.3.1 *Using hybrid blocklists.*

Creating hybrid blocklists – lists that combine algorithmic and social curation (Section 5.1) – can be a promising approach. Such lists could rely on carefully constructed algorithms that surface offensive content and categorize it based on severity. Posters of blatantly abusive content can be blocked directly. For postings that are flagged by such algorithms as possibly abusive, human moderators can examine them and decide whether the posters should be blocked. These blocklists should also have sufficient fail-safe mechanisms built into them so that the actions of an intruder or a rogue moderator may be quickly reverted (Section 5.4).

Our findings indicate that some users found themselves on popular blocklists because of a tenuous connection with controversial individuals on Twitter (Section 4.1). Human moderators can ensure that individuals are not blocked for trivial reasons like following an abusive user. They may also consider muting certain individuals instead of blocking them so as to avoid punishing certain actions disproportionately.

Such a hybrid mechanism would make the curation of blocklists more objective and efficient as well as ameliorate the risks of having a large number of false positives (Section 5.5). This mechanism could also be adapted to improve the accuracy of blocklists that are curated for purposes other than addressing harassment, e.g., spam blocklists.

6.3.2 *Making blocklists more transparent.*

Our findings in Section 5.5 show how branding a blocklist as a list of harassers can be dangerous, particularly if the list contains many false positives. The blocklist owners have a responsibility to communicate to their subscribers that some users may mistakenly be on the list. They should also make efforts to ensure that the users on the list are not discriminated against. Participant SB-11 described a few such efforts she made for Block Bot:

> *"Back in the day... the way that things were written out on their [Block Bot's] website were more blunt... I changed a lot of the wording so that it was a little less harsh. That seemed to help people not be as upset about it. We've always been a little irreverent about complaints because all it is is blocking someone, we're not saying that you're a terrible person or that whatever you do on Twitter rises to some legal definition of harassment or anything like that."- SB-11*

This suggests that clarifying the purpose of the blocklist can help de-escalate rancor from the blocked users. The blocklist administrators can also choose to explicitly discourage discrimination against blocked users for any purpose outside Twitter.

Different user groups have different definitions of harassment and distinct moderation needs (Section 5.4), and therefore they may need to subscribe to different blocklists. Therefore, multiple instances of blocklists should be constructed. Each such instance should clearly state its purpose, and its moderators should be aware of and have the capability to address the needs of its subscribers. Moderators should be encouraged to reveal aspects of their identities and experiences that shape how they moderate. This would allow subscribers to take into account the biases of the moderators before subscribing to any blocklist. We found that many users who were put on blocklists were frustrated because they did not know the reason for their being put on the list (Section 5.6). We posit that blocklists should record the reason why each account is blocked, the moderator who blocked it, and other relevant metadata. Providing information about the reason for being put on the blocklist when requested may encourage the acceptance of blocklists among many users.

6.3.3 *Designing to avoid blocking contagion.*

In her book on community self-regulation, Ostrom writes that "graduated punishments ranging from insignificant fines all the way to banishment, applied in settings in which the sanctioners know a great deal about the personal circumstances of the other appropriators and the potential harm that could be created by excessive sanctions, may be far more effective than a major fine imposed on a first offender" [35]. Drawing from Ostrom's work, Kiesler et al recommend using graduated sanctions to increase the legitimacy and effectiveness of sanctions in online communities [25]. They argue that "lighter sanctions mitigate the ill effects from inevitable mistakes in categorization" and "stronger sanctions are perceived as more legitimate when applied only after lighter sanctions have proven ineffective." In a similar vein, Forte and Bruckman found that in order to maintain local standards of content production, Wikipedia uses a series of graduated sanctions when behavior-related policy is broken – beginning with the posting of warnings and leading to banning from the site [17].

We discussed in Section 5.5 that blocking contagion could be a serious consequence of the popular use of Twitter blocklists. To prevent this problem, platforms like Block Together can draw lessons of graduated sanctions from the research described above and discourage the permanent blocking of blocked accounts. Instead, they can consider enforcing the blocks only for a limited time interval initially, and escalate sanctions if repeated misbehavior occurs, for example, by increasing the time interval for which the offending user is blocked.

Blocking mechanisms can also consider discouraging the outright copying of blocked accounts for creation of new blocklists. They can make such copying contingent upon the permission provided by some central moderators. A central supervision of blocklists that are currently in use, and a regular evaluation of whether they serve their stated purpose, would guard against misappropriation of blocklists.

6.3.4 Improving appeals procedure. We discussed in Section 5.7 that a dissatisfactory appeals procedure delegitimized the use of blocklists for many participants. The process of appealing to get oneself off any blocklist should be made more intuitive and efficient. Timely and appropriate responses to such appeals, along with an effort to spread awareness about the damaging effects of online abuse, would help such blocking mechanisms gain broader popularity on the site. We acknowledge that responding to appeals is expensive for the blocklist administrators. Therefore, we recommend focusing on having fewer false positives, and automating the process of responding to certain types of appeals.

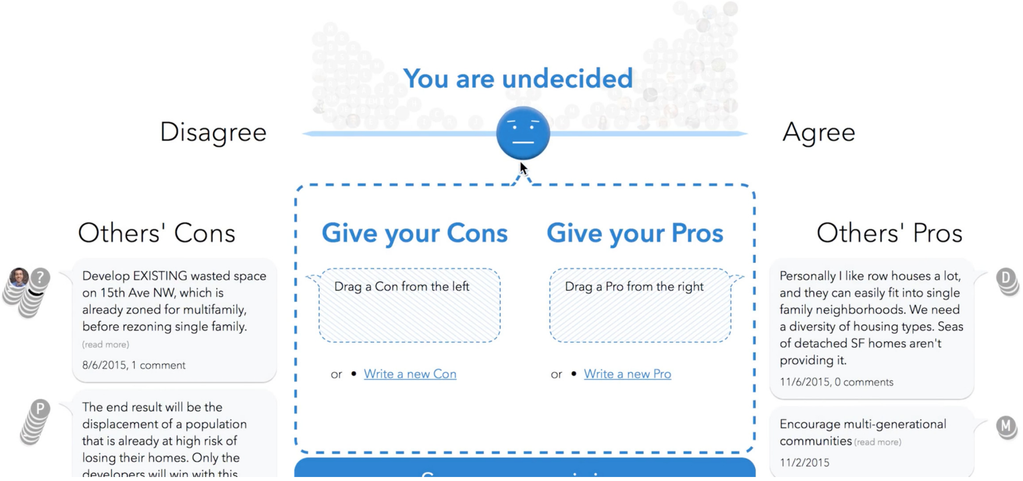
6.4 Building "understanding mechanisms"

Our findings indicate that differences in identities, perspectives and sensibilities often contribute to situtations where some users perceive that they are being harassed and other users see it as mere disagreements (Section 4.1). Additionally, (mis)interpreting the words of the opposite side in a negative light and reacting inordinately over incidents of minor disagreements create further rifts and preclude productive discourse. Differences in behavioral and cultural norms across different user groups further escalate such situations. Furthermore, as Van Alstyne and Brynjolfsson warned in their study on "cyberbalkanization," if people spend more time on special interests and screen out less preferred content, it can "fragment society and balkanize interactions" [49].

To address these challenges, designers need to focus on creating "understanding mechanisms." Tools that emphasize similarities between individuals could help them to understand one another and find common ground. Design solutions that allow users with different ideologies to interact without fear of being abused could foster productive discussions. There is a growing body of literature on modeling argumentation for the social semantic web [41]. Designers can draw from the theoretical models and social web tools that argumentation researchers have proposed to implement mechanisms that facilitate constructive discussions.

Consider an open-source deliberation platform developed at the University of Washington, ConsiderIt [26], that powers the Living Voters' Guide[16]. This platform invites users to think about

[16] https://livingvotersguide.org



Develop EXISTING wasted space on 15th Ave NW, which is already zoned for multifamily, before rezoning single family. (read more)
8/6/2015, 1 comment

The end result will be the displacement of a population that is already at high risk of losing their homes. Only the developers will win with this

Personally I like row houses a lot, and they can easily fit into single family neighborhoods. We need a diversity of housing types. Seas of detached SF homes aren't providing it.
11/6/2015, 0 comments

Encourage multi-generational communities (read more)
11/2/2015

Fig. 4. ConsiderIt allowing a user to compile pro and con lists for this proposal for Seattle city: “Increase the diversity of housing types in lower density residential zones.”

the tradeoffs of a proposed action by creating a pro/con list (Figure 4). This list creation is augmented by allowing users to include into their own list the points that have already been contributed by others. This process allows users to gain insights into the considerations of people with different perspectives and identify unexpected common ground [26]. Additionally, the platform’s focus on personal deliberation, as opposed to direct discussion with others, reduces the opportunities for conflicts [26].

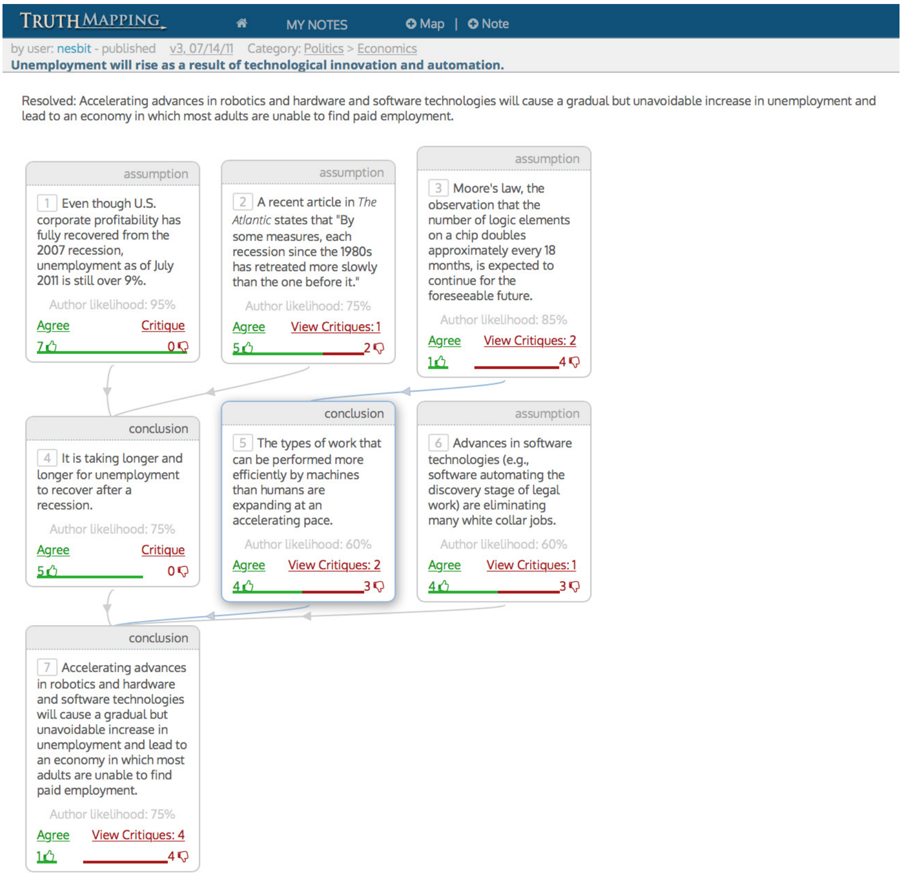
TruthMapping [17] is another online deliberation tool that allows users to collect and organize ideas, constructively test those ideas, and promote reasoning-based discourse. This tool structures conversations using argument maps, critiques and rebuttals (Figure 5). It invites users to break down a topic into its component parts – assumptions and conclusions – and create a node for each part, so that the hidden assumptions are made explicit. All critiques are directed against specific nodes so that any attempts at digression are apparent. Only the original arguer can modify the map but any user can provide their feedback by adding a critique to any assumption or conclusion or by responding to a previously posted critique with a rebuttal. As shown in Figure 5, TruthMapping also shows how many users agree or disagree with each node.

Although designs like ConsiderIt and TruthMapping offer innovative solutions to facilitating constructive deliberation, they assume that the users are participating in good faith, and are willing to devote their time to review previously posted content and submit productive accessions. These assumptions may not be true for many participants on social media sites. Therefore, designing “understanding mechanisms” for SNSs is a considerably hard problem, and there is a lot of potential for researchers to experiment with creative solutions in this space.

7 CONCLUSION

Online harassment is a multi-faceted problem with no easy solutions. Social media websites are persistently squeezed between charges of indifference to harassment and suppression of free speech [3]. We believe it is an important and difficult challenge to design technical features of SNSs and

[17] https://www.truthmapping.com



by user: nesbit - published v3, 07/14/11 Category: Politics > Economics

Unemployment will rise as a result of technological innovation and automation.

Resolved: Accelerating advances in robotics and hardware and software technologies will cause a gradual but unavoidable increase in unemployment and lead to an economy in which most adults are unable to find paid employment.

assumption
1 Even though U.S. corporate profitability has fully recovered from the 2007 recession, unemployment as of July 2011 is still over 9%.
Author likelihood: 95%
Agree Critique
7 0

assumption
2 A recent article in *The Atlantic* states that "By some measures, each recession since the 1980s has retreated more slowly than the one before it."
Author likelihood: 75%
Agree View Critiques: 1
5 2

assumption
3 Moore's law, the observation that the number of logic elements on a chip doubles approximately every 18 months, is expected to continue for the foreseeable future.
Author likelihood: 85%
Agree View Critiques: 2
1 4

conclusion
4 It is taking longer and longer for unemployment to recover after a recession.
Author likelihood: 75%
Agree Critique
5 0

conclusion
5 The types of work that can be performed more efficiently by machines than humans are expanding at an accelerating pace.
Author likelihood: 60%
Agree View Critiques: 2
4 3

assumption
6 Advances in software technologies (e.g., software automating the discovery stage of legal work) are eliminating many white collar jobs.
Author likelihood: 60%
Agree View Critiques: 1
4 3

conclusion
7 Accelerating advances in robotics and hardware and software technologies will cause a gradual but unavoidable increase in unemployment and lead to an economy in which most adults are unable to find paid employment.
Author likelihood: 75%
Agree View Critiques: 4
1 4

Fig. 5. TruthMapping allows users to construct an argument by laying out assumptions and conclusions.

seed their social practices in a way that promotes constructive discussions and discourages abusive behavior.

The emergence of third-party, open-source moderation mechanisms like blocklists introduces an innovative alternative to traditional centralized and distributed moderation systems. In this study, we focused on studying the effects of using blocklists - on those who used them and those who were blocked on them. We also used blocklists as a vehicle to investigate the broader issue of online harassment.

This paper does not investigate all the possible forms and aspects of online harassment. Participants in the study were strategically recruited in ways that ensured awareness of and experience with these issues on Twitter. Other methods of recruiting may reveal other, perhaps more commonplace, experiences of average users with undesirable content and moderation. Researchers may also recruit users from specific vulnerable groups to understand their particular experiences and needs. Do these groups need moderation tools that serve their special needs? Can we design to detect distinctive harassment strategies such as dogpiling and brigading? Can we construct tools to combat these strategies that are not vulnerable to being abused? We continue to pursue these questions in our ongoing investigations of online harassment.

In the interim, by describing the experiences of users affected by blocklists on Twitter, we see concrete examples of the gap between the needs of users and the affordances provided by default and third-party moderation mechanisms on social media. If we hope to create scientifically informed guidelines for designers to follow, more work is needed that tests innovative design ideas for improved moderation in lab and field experiments.

ACKNOWLEDGMENTS

We are grateful to our participants for sharing their experiences on sensitive issues like online harassment and content moderation. We would also like to thank Stevie Chancellor, Michaelanne Dye, Julia Deeb-Swihart and three anonymous reviewers who offered generous feedback on earlier versions of this work.

REFERENCES

[1] Zahra Ashktorab and Jessica Vitak. 2016. Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, New York, New York, USA, 3895–3905.

[2] Associated Press. 2017. Unreal when it targets you: Faceless trolls attack online. (2017). http://molawyersmedia.com/2017/04/14/unreal-when-it-targets-you-faceless-trolls-attack-online-2/

[3] David Auerbach. 2016. If Only AI Could Save Us from Ourselves. (2016). https://www.technologyreview.com/s/603072/if-only-ai-could-save-us-from-ourselves/

[4] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

[5] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2018. Classification and its Consequences for Online Harassment: Design Insights from HeartMob.. In *Proceedings of ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '18 Online First).*

[6] The BlockBot. 2016. The Block Bot. (2016). http://www.theblockbot.com

[7] danah boyd. 2008. Why Youth Heart Social Network Sites: The Role of Networked Publics in Teenage Social Life. *MacArthur Foundation Series on Digital Learning – Youth, Identity, and Digital Media* (2008), 119–142. DOI: http://dx.doi.org/10.1162/dmal.9780262524834.119

[8] Kathy Charmaz. 2006. *Constructing grounded theory: a practical guide through qualitative analysis.* London. DOI: http://dx.doi.org/10.1016/j.lisr.2007.11.003 arXiv:arXiv:1011.1669v3

[9] Danielle Keats Citron. 2014. *Hate crimes in cyberspace.* Harvard University Press.

[10] Robyn M Cooper and Warren J Blumenfeld. 2012. Responses to Cyberbullying: A Descriptive Analysis of the Frequency of and Impact on LGBT and Allied Youth. *Journal of LGBT Youth* 9, 2 (2012).

[11] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (mar 2016), 410–428. DOI: http://dx.doi.org/10.1177/1461444814543163

[12] Munmun De Choudhury, Shagun Jhaver, Benjamin Sugar, and Ingmar Weber. 2016. Social Media Participation in an Activist Movement for Racial Equality. In *Tenth International AAAI Conference on Web and Social Media.*

[13] Jill P. Dimond, Michaelanne Dye, Daphne Larose, and Amy S. Bruckman. 2013. Hollaback! the role of storytelling online in a social movement organization. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work - CSCW '13.* ACM Press, New York, New York, USA, 477. DOI: http://dx.doi.org/10.1145/2441776.2441831

[14] Judith S Donath. 1999. Identity and deception in the virtual community. *Communities in cyberspace* 1996 (1999), 29–59.

[15] Maeve Duggan. 2014. Online Harassment. *Pew Internet Project* (2014).

[16] Maeve Duggan. 2017. Online Harassment. *Pew Internet Project* (2017).

[17] Andrea Forte and Amy Bruckman. 2008. Scaling Consensus: Increasing Decentralization in Wikipedia Governance. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008).* IEEE, 157–157. DOI: http://dx.doi.org/10.1109/HICSS.2008.383

[18] GamerGate Wiki. 2016. The Block Bot. (2016). http://thisisvideogames.com/gamergatewiki/index.php?title=The

[19] GamerGate Wiki. 2017. GGAutoBlocker - GamerGate Wiki. (2017). http://thisisvideogames.com/gamergatewiki/index.php?title=GGAutoBlocker

[20] R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016).

[21] Ralph Gross and Alessandro Acquisti. 2005. Information revelation and privacy in online social networks. *Proceedings of the 2005 ACM workshop on Privacy in the electronic society* (2005), 71–80. DOI: http://dx.doi.org/10.1145/1102199.1102214

[22] Amanda Hess. 2014. Twitter harassment: User-created apps Block Together, Flaminga, and the Block Bot crack down on Twitter abuse. (2014). http://www.slate.com/blogs/future

[23] Zachary Jason. 2015. Game of Fear. (2015). http://www.bostonmagazine.com/news/article/2015/04/28/gamergate/

[24] Shagun Jhaver, Larry Chan, and Amy Bruckman. 2017. The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action. *Under review* (2017).

[25] Sara Kiesler, Robert Kraut, and Paul Resnick. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design* (2012).

[26] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*. ACM Press, Seattle, Washington, USA, 265–274. http://dl.acm.org/citation.cfm?doid=2145204.2145249

[27] Cliff Lampe and Paul Resnick. 2004. Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*.

[28] Amanda Lenhart, Michele Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. 2016. Online Harassment, Digital Abuse, and Cyberstalking in America. (2016). https://datasociety.net/output/online-harassment-digital-abuse-cyberstalking/

[29] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and Others. 2011. The Arab Spring - the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International journal of communication* 5 (2011).

[30] David Malki. 2014. The Terrible Sea Lion. (2014). http://wondermark.com/1k62/

[31] J Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. 2015. Reporting, reviewing, and responding to harassment on Twitter. (2015).

[32] Sharan B Merriam. 2002. Introduction to Qualitative Research. *Qualitative research in practice: Examples for discussion and analysis* 1 (2002).

[33] Elise Moreau. 2016. The Top 25 Social Networking Sites People Are Using. (2016). https://www.lifewire.com/top-social-networking-sites-people-are-using-3486554

[34] Torill Elvira Mortensen. 2016. Anger, Fear, and Games The Long Event of #GamerGate. *Games and Culture* (2016).

[35] Elinor Ostrom. 1990. Governing the commons: the evolution of institutions for collective action. (1990).

[36] Patreon. 2017. Randi Harper is creating Online Activism and Open Source Anti-Harassment Tools — Patreon. (2017). https://www.patreon.com/freebsdgirl

[37] Whitney Phillips. 2015. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture.* MIT Press.

[38] Dan Pine. 2016. Anti-Semitic emails, tweets hit candidates and journalists. (2016). http://www.jweekly.com/2016/06/17/anti-semitic-emails-tweets-hit-candidates-and-journalists/

[39] RationalWiki. 2016. Atheism Plus. (2016). http://rationalwiki.org/wiki/Atheism

[40] Rebecca Ruiz. 2014. When Your Job Is to Moderate the Internet's Nastiest Trolls. (2014). http://mashable.com/2014/09/28/moderating-the-trolls/

[41] Jodi Schneider, Tudor Groza, and Alexandre Passant. 2013. A review of argumentation for the social semantic web. *Semantic Web* 4, 2 (2013), 159–218.

[42] Sjwomble. 2016. The Problem With Personal Block Lists. (2016). https://sjwomble.wordpress.com/2016/04/28/the-problem-with-personal-block-lists/

[43] Todd Spangler. 2017. IMDb Shuts Down Discussion Boards — Variety. (2017). http://variety.com/2017/digital/news/imdb-message-boards-shut-down-1201977581/

[44] Statista. 2017. Twitter: number of active users 2010-2017. (2017). http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[45] Nitasha Tiku and Casey Newton. 2015. Twitter CEO: 'We suck at dealing with abuse' - The Verge. (2015). http://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the

[46] S Turkle. 2006. Life on the screen: Identity in the age of the internet. (2006). http://www.citeulike.org/group/48/article/949801

[47] Twitter. 2016. Blocking accounts on Twitter. (2016). https://support.twitter.com/articles/117063

[48] Twitter. 2016. What are replies and mentions? (2016). https://support.twitter.com/articles/14023

[49] Marshall van Alstyne and Erik Brynjolfsson. 1996. Electronic Communities: Global Villages or Cyberbalkanization? *ICIS 1996 Proceedings* (1996). http://aisel.aisnet.org/icis1996/5

[50] Theresa Velden and Carl Lagoze. 2013. The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication. *Journal of the American Society for Information Science and Technology* (2013).

[51] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '17*. DOI: http://dx.doi.org/10.1145/2998181.2998337

[52] Kyle Wagner. 2012. The Worst Job at Google: A Year of Watching Beastiality, Child Pornography, and Other Terrible Internet Things. (2012). http://gizmodo.com/5936572/the-worst-job-at-google-a-year-of-watching-beastiality-child-pornography-and-other-terrible-internet-things

[53] Charlie Warzel. 2016. "A Honeypot For Assholes": Inside Twitter's 10-Year Failure To Stop Harassment. (2016). https://www.buzzfeed.com/charliewarzel/a-honeypot-for-assholes-inside-twitters-10-year-failure-to-s

Received February 2017; Revised October 2017; Accepted January 2018