# After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities



JIE CAI, New Jersey Institute of Technology, USA
DONGHEE YVETTE WOHN, New Jersey Institute of Technology, USA

Content moderation is an essential part of online community health and governance. While much of extant research is centered on what happens to the content, moderation also involves the management of violators. This study focuses on how moderators (mods) make decisions about their actions after the violation takes place but before the sanction by examining how they "profile" the violators. Through observations and interviews with volunteer mods on Twitch, we found that mods engage in a complex process of collaborative evidence collection and profile violators into different categories to decide the type and extent of punishment. Mods consider violators' characteristics as well as behavioral history and violation context before taking moderation action. The main purpose of the profiling was to avoid excessive punishment and aim to integrate violators more into the community. We discuss the contributions of profiling to moderation practice and suggest design mechanisms to facilitate mods' profiling processes.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: live streaming; content moderation; volunteer moderator; profiling

**ACM Reference Format:**
Jie Cai and Donghee Yvette Wohn. 2021. After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 410 (October 2021), 25 pages. https://doi.org/10.1145/3479554

## 1 INTRODUCTION

Online abuse, such as hateful speech, sexual harassment, personal attack, and doxing, is a severe and pervasive social problem that many. According to a research survey in January 2020, 44% of Americans report that they experience online harassment. In some cases, these experiences are coupled with other impacts, such as anxiety and thoughts of depression and suicide [1].

In order to reduce online abuse and maintain the growth and health of online communities, commercial platforms apply many techniques to filter abusive language, such as improving algorithms and applying tools (e.g., [2, 11, 43, 45]). However, violators always seek ways to circumvent the algorithms and cheat the tools with variants [10, 30]. To supplement algorithmic moderation, platforms also rely on human moderators (mods), either active volunteer users [66] or well-trained content experts [58], to manually remove user-reported content or review incidents in context-sensitive situations [62]. When dealing with harmful content, mods also try not to push hard to

Authors' addresses: Jie Cai, New Jersey Institute of Technology, Newark, USA, jie.cai@njit.edu; Donghee Yvette Wohn, New Jersey Institute of Technology, Newark, USA, yvettewohn@gmail.com.


2573-0142/2021/10-ART410 $15.00
https://doi.org/10.1145/3479554

alienate community members [69] as many presumed violators feel frustrated and complain that the content removal is unfair and lacks transparency [42, 44, 51].

Since community growth and health is about not only punishing but also maintaining users by setting positive examples [65], understanding users' characteristics is a good way to avoid sanctioning users by mistake and to improve the perceived justice and fairness. Checking a user's account information, which is a good indicator of a user's characteristics (e.g., [27, 37, 74]) and a reference to the user's commonalities with others (e.g., [52, 68]), is one way to do so. As account information and activities reveal users' behaviors, profiling, which refers to the dynamic process of collecting and integrating users' information and activities to find their behavioral patterns and characteristics [49], is vital for mods to understand bad actors, a challenge highlighted by prior research [42, 45].

In line with recent HCI and CSCW research proposing to understand bad actors [4, 48], this work aims to explore how mods psychologically profile violators in live streaming communities. Due to the lack of HCI theories associated with profiling, we used criminal profiling [40] as a lens to understand moderators' mental models about the profiling process and types of violators. To achieve this goal, we first observed moderation work through mods' self-recorded videos. Using videos as probes, we then interviewed mods while watching the videos. In the interviews, we asked their reasoning to deal with violators, such as the information they are looking for and the reason for their judgment.

This work contributes to understanding mods' mental models regarding what happens after violation but before sanction. In most cases, profiling allows mods in micro-communities to understand violators' characteristics to avoid excessive punishment or, more importantly, mediate and support community members. We present their profiling process, how they collect information for profiling, and the violators' types with various moderation strategies. We discuss how the platform's affordances and design affect profiling ; we also discuss how profiling can potentially grow the community through increasing justice and fairness and distinguishing bad actors. Finally, we suggest social and technical interventions that could assist in profiling in the moderation process.

## 2 BACKGROUND AND RELATED WORK

### 2.1 Content Moderation

Moderation refers to *"the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse"* [33] and is the gateway for online communities to thrive as harassment, trolls, and hate speech are increasing in these spaces [4], ranging broadly from learning communities (e.g., [67, 81]) to crowd-sourcing communities (e.g., [14, 19]) to social media platforms (e.g., [20, 45, 58, 66]). Moderators (mods) are gatekeepers [58] of commercial platforms to maintain the community health and growth [69] with the power to remove harmful content and sanction users posting the content, namely violators. However, abusing moderation power or overly sanctioning users could deter community engagement and alienate community members [69]. Mods have to trade off the punishment efficacy and community growth [14, 69]. Mods in community moderation play various fluid roles to shape the communities [64]; they collaborate with other mods [55] and apply moderation tools to curate content [43], and help the community leader to manage user engagement [80]. Mods also suffer from emotional tolls like lack of appreciation from the community administrator [79] and have to handle the emotional labor [22]. They might even experience impairment of psychological well-being [70] due to facing harmful content and doing the dirty work for a long time [58]. Though a lot of research explores how mods handle violations and their impact, there is a lack of understanding of how mods perceive violators/bad actors.

According to the moderation sequence, prior moderation work focuses on 1) proactively preventing mechanism, 2) reactively sanctioning mechanism, 3) the impact of sanctioning on users. A thread of research focuses on proactively preventing bad acts through norm-setting [13, 26] and increasing the violation cost [33]. For example, Seering et al. [63] reveals that using small interventions like CAPTCHAs as stimuli to promote a positive mindset of users can successfully activate users' positive affect and their behaviors in the chat. Another thread of research focuses on reactively sanctioning users, such as deleting content and banning users. This thread of research mainly explores how mods collaborate with moderation tools [43] and how to use computational approaches to automatically detect and filter harmful content, such as labeling and classification [3] and moderation tool design [5, 11]. The third thread of research focuses on the impact of sanctions on users, either normal users [45, 65] or violators [14, 42, 44]. For example, Seering et al. [65] finds that banning a particular type of behavior decreases the frequency of that behavior performed by normal users in the following messages in Twitch chat. On Wikipedia, nearly half of first-time violators who are temporarily blocked either recidivate or abandon the community [14].

While most prior work investigates the prevention of violation, the sanction mechanism, and the impact of sanction on users, there is a lack of understanding of what happens after the violation but before a sanction, a small gap in our understanding of the moderation sequence. Aside from removing content and sanctioning users, recent work shows that educating violators is an effective strategy to get rid of toxicity [6]. Others suggest that when considering online behaviors as harassment, we should also consider the context [9]. We are motivated by such results in moderation work and explore the mental process of mods. Why do mods decide to punish some but not others behaving similarly? How do they decide who should be punished?

### 2.2 Criminal Profiling

Criminal profiling, also called "psychological profiling" or "offender profiling", is "an educated attempt to provide investigative agencies with specific information as to the type of individual who committed a certain crime" [29]. Similarly, Egger [25] defines criminal profiling as "an attempt to provide investigators with more information on the offender who is yet to be identified." Generally, criminal profiling is the process of gathering evidence both at the scene of a crime and from the victims and witnesses to construct a biographical sketch of the criminal [50]. Hicks and Sales [40], in their book dedicated to the development of criminal profiling, propose that crime scene evidence is the primary source of investigative information available to investigators, including physical evidence and victim information and statements, and that the offender's characteristics cause them to leave particular pieces and patterns of evidence during the crime. Through these shreds of evidence, the investigator pieces together the offender's characteristics to figure out the types of offenders.

Focusing on the roles that evidence can play in informing a timeline and narrative of the crime, Chisum and Rynearson [16] classify physical evidence into different types such as sequential (sequence of events surrounding a criminal act), directional (where something was going and coming from), location (position and orientation of people and objects surrounding the scene), and limiting (the nature and boundaries of the crime scene) evidence. The breakdown of evidence can facilitate answering "who," "what," " when," "where," "how," and sometimes "why" questions about the commission of the crime [17]. The evidence is usually collected by the crime scene investigation team consisting of photographers and specialists and then sent to forensic psychologists and other experts to analyze [57]. Criminal profiling shows how a group of researchers systematically collect evidence and deduct criminals' characteristics based on the evidence and is worthwhile for police investigation because of its improvement in the scientific rigor of research (e.g., see the meta-analysis by [23, 28]).

Criminal profiling plays different roles in the criminal justice system in three phases: criminal investigation, apprehension, and prosecution [40]. In the investigation phase, profiling aims to link evidence as part of a series to identify physical and psychological characteristics of unknown offenders, to predict the pre-and post-offense behaviors that an offender might show, and to evaluate the potential escalation of certain criminal behaviors. In the apprehension phase, profiling suggests evidence collection on the search warrants or interrogation techniques eliciting a confession from an offender and predicts an offender's behaviors on the arrest. In the prosecution phase, profiling works as providing expertise in the courtroom to demonstrate the linking of multiple offenses to one individual or to match a particular individual to the relevant crime(s) [40, p13]. In this study, we mainly focus on the investigation phase, which aims to understand violators' characteristics or evaluate violators' behaviors to avoid similar violations happening in the future.

### 2.3 Applying Criminal Profiling to Community Moderation and in Live Streaming Communities

Much research in HCI discusses profiling normal users online, such as how to develop different types of clustering, how to use algorithms to cluster users and develop different personas (e.g., [15, 24, 59]), and how to predict users' preferences and provide better services (e.g., [34, 56, 72, 73]). Stainbock [71] reveals the connection of general profiling using algorithms and criminal profiling and states that "data mining's computerized sifting of personal characteristics and behaviours (sometimes called 'pattern matching') is a more thorough, regular, and extensive version of criminal profiling". In these contexts, the person conducting the profiling is usually an industry professional and targets the regular user but not violators. Little research in user profiling literature focuses on collecting the moderated information to profile violators, the information that is removed and invisible to the public. Mods have access to both the normal content visible to the public and the invisible violative content, owning the advantage to see the holistic scenario to understand a user's behavior and characteristics. Though some work focuses on collecting moderated information to understand violations, no specific work applies the profiling lens to understand violators.

In community moderation, criminal profiling has been used as a lens to exemplify how Wikipedia moderation tools work as profiling agents, from observing and catching vandalistic edits to finally generating patterns using either structured decision-making or a black-box approach [21]. Additionally, some research points out the necessity for human mods to collect evidence for their decision-making during the moderation process. For example, Jiang et al. [46] found in live voice chat on Discord, mods face challenges to collect evidence of potential violators, sometimes even with the risk of violating privacy policy to secretly record voice as evidence. Kiene et al. [47] also found that the moderation tools are insufficient for organizing and retrieving information for mods to make consistent decisions towards violations and that mods seek user-developed bots to track information of community members. Research in live streaming communities shows that some mods use moderation tools to check a viewer's history [7] but are not satisfied with the features of these tools and hope to have more information about viewers and violators [5]. While this thread of research discusses the need and necessity of more evidence for moderation, they do not specify what type of evidence they need, how they collect the evidence, and consequently, how to use these shreds of evidence to evaluate potential violations and punish potential violators.

The need of understanding evidence collection in community moderation and the lack of framework in user profiling literature to understand violators indicate the potential of a new lens to build a connection between community moderation and profiling research. Additionally, much research also introduces various types of justice (e.g., social justice, retributive justice, restorative justice) from the criminal justice system to explain online harassment and moderation, and the justice-seeking process [2, 21, 60, 61, 70]. The inherent role of criminal profiling in the criminal

justice system and its components (evidence collection and analysis, and the deduction of offender's characteristics) in the definition suggests that criminal profiling may serve as a good lens to understand mods' profiling process when they face potential violators at a conceptual level. In line with the definition of criminal profiling and applying it to live streaming communities, we asked the following questions:

- **RQ1:** *What kind of evidence do mods collect to profile violators?*
- **RQ2:** *How do mods collect these types of evidence?*
- **RQ3:** *What are the types of violators that mods perceive?*

The online environment makes the application of this lens slightly different from the offline world. First, online behaviors become part of the evidence. In online communities, harmful content is considered crime scene evidence and reflects online behavior. Most types of evidence relevant to physical evidence (e.g., blood, body drag, glass fragments) are not applicable to the online environment. Second, the offender is already identified in live streaming communities, so the profiling is not to find the offender but to evaluate whether or not the mod should punish him and to what extent the punishment should be. Instead of directly banning users, they may also look for other evidence. Third, researchers in criminal profiling rely heavily on the captured offenders' self-reported information to figure out their characteristics. In live streaming communities, mods as non-experts directly communicate with violators and can access various information.

## 3 TWITCH AS THE RESEARCH SITE

As a unique social medium with high-fidelity computer graphics and video and low-fidelity text-based communication [36], live streaming is a rapidly growing industry. Twitch has become a global leading live steaming platform, starting from gaming content and expanding into a range of all imaginable content categories. In early 2020, it had more than 3 million active monthly creators and over 15 million average daily streamers [1]. It is estimated to surpass 40 million US users by the end of 2021 [2]. On Twitch, users can create their profiles under the profile settings, such as updating profile picture (displayed as a head image), adding profile banner (displayed as the background on the top of their homepages), changing username (username updates can be performed once every 60 days), and adding bio information (displayed as "About" if other users check their profile). When joining a chatroom, users can click a viewer's username to see the viewer's basic information in the channel. A further click of the username will forward the user to the viewer's homepage.

To handle the user-generated content, Twitch employs a multi-layered moderation system, including both automated moderation tools and human mods, although it continues to change its structure. At a broad level, the company has employees who are well-trained people and mostly handle inappropriate broadcasting content that has been reported by users with common criteria for the entire community [75]. At a micro level, Twitch users form micro-communities [79] around streamers, and streamers appoint volunteer mods who are active community members to handle other users and messages in the chat with specific criteria. Since each micro-community operates under different criteria, users may behave variously across different micro-communities. Also, streamers and mods can choose to activate/deactivate a moderation tool called AutoMod that uses algorithms to filter abusive messages. Twitch also has an open-access API for the integration of thirty-party moderation tools. Mods have to track a high volume of fast-moving messages, identify the negative ones, and take action within a limited time because of the nearly-synchronous conversation in the chat [66, 79]. This poses unique challenges because it means they have very limited time to make decisions about what moderation actions they will take. The interactive social

[1] https://www.twitch.tv/p/press-center/
[2] https://www.emarketer.com/newsroom/index.php/twitch-on-pace-to-surpass-40-million-viewers-by-2021/

medium context with unique challenges of moderation, in addition to the lack of understanding of how human mods profile violators in community moderation, motivated us to focus on volunteer mods and their moderation of viewers and messages in the chat.

## 4 METHODS

Since the profiling process happens behind the scene, we chose the observation plus interview method to explore the research questions. The observation allowed us to see the whole moderation process, from seeing a violation to finally sanctioning the violator. The interview alongside the video allowed us to recall the moderation actions with mods and then to ask questions about their decision-making process. The school IRB approved this project, and the consent form was sent to participants before the interview through either email or Discord.

We offered two options for mods to participate. The first option (A) was to share with us a self-recorded video of the screen when they were moderating. After we reviewed the video, we scheduled the interview. Mods received a $100 Amazon gift card after the interview. Because some mods had strong privacy and safety concerns and/or felt uncomfortable with recording, we provided a second option (B) with a $50 Amazon gift card, only conducting a semi-structured interview but asking them to provide necessary examples (e.g., screenshots, video clips) during the interview.

### 4.1 Participant Recruitment and Demographics

We recruited 19 participants through three approaches. First, we reached the potential participants through the email list that we collected from Twitch Convention 2019 and received six responses. Twitch Convention is a gathering of the Twitch community hosted by Twitch to provide the opportunity for streamers, moderators, viewers, and merchandisers to meet offline. We recruited from Twitch Convention offline to increase diversity and avoid the bias of only recruiting people online. Second, one research assistant who was also a Twitch mod asked other mods in the channels he moderated to recruit five participants. Third, we used our personal Twitch accounts and browsed the recommended channels on the Twitch homepage. We first entered live channels to observe for 5-10 minutes. After we saw active mods, we asked and obtained eight mods. We had 12 male mods and seven female mods. The average age was 23. Most mods were white. The average moderation experience was three years, ranging from half a year to eight years. Most primarily moderated gaming communities. 10 mods chose option A, and nine mods chose option B. The viewership of the channel in option A varied from tens to thousands. Details are summarized in Appendix A.

### 4.2 Video Analysis and Interview Process

We first ran a pilot study with the mod in our team. Three researchers interviewed the mod to test the flow of the interview protocol and watched the mod's moderation practices to decide the reasonable length of recorded video for the observation. The pilot video was one and a half hours long, with 105 active viewers on average in the chat. We observed a lot of violations and repeating moderation in the full video, even in the first hour, and thus considered one hour a reasonable length for the observation. All the participants were encouraged to share with us a one-hour-length video through Google Drive. To analyze the video, we focused on moderation related actions and developed a codebook for video coding (1, explain; 2, delete; 3, warning; 4, timeout; 5, ban; 6, should have moderated but not (ignored); 0, other interesting issues). An explain is the rule explanation in the chat; a delete means the message was removed in the chat; a warning means sending a warning message to the viewer in the chat; a timeout is a temporary block from minutes to hours; a ban is a permanent block, indicating the violator can not send a message in the chat anymore. Warning, delete, timeout, and ban can be achieved via bot command that is alongside the username and badges, as shown in Figure 1. In the coding process, we focused on these actions and excluded

mods' social interaction, such as greeting newcomers and just chatting with viewers. Each video was analyzed by two researchers separately to identify the timestamps of each relevant action. Then, the two researchers discussed their results to achieve consistent timestamps.

All the interviews were conducted after the video analysis and through Discord. Before the interview, we opened the recorded video on our side and also asked mods to open the video on their side. In the interview, we first asked some general questions about their moderation experience, such as which platforms they moderate for and how long they have been moderating. Then we asked some questions about providing examples of moderation decisions they made. Later, we asked them to look at the video for each timestamp that we noted and to explain their decisions. For example, "At 35:04 (35 mins and four secs), I saw you deleted the message and banned the user. What was your rationale to make that decision?" For mods who chose option B, we skipped these questions. After questions on video analysis, we asked questions about profiling, such as what kind of information helped them moderate and what the reasons/motivations were for users to perform badly. Since option B did not share video to help us gain context, we often asked follow-up questions such as "do you have a specific example to show us?", "can you give us an example?" and "can you explain more about this?" These follow-up questions reminded them of something they recorded and saved from their end. They thus shared the content with us via Discord during the interview. In the end, we asked for demographic information. The interview protocol and process followed a consistent structure for both options, except that the video plus interview option added several questions for each timestamp, and that the interview-only option asked more follow-up questions about examples. All interviews were audio-recorded, transcribed by speech recognition software [3], and then double-checked by the researchers.

### 4.3 Interview Analysis

We imported all transcripts into ATLAS.ti Cloud [4] for collaborative coding. First, four researchers individually went through all transcripts to have general ideas. Next, four researchers picked up a transcript with abundant content to code individually. After individual coding, four researchers had a group meeting to discuss the codes and clarify the definitions. All codes with definitions were archived. Four researchers, repeating the above steps, coded three transcripts to develop an initial codebook. By following the initial codebook, each transcript of the rest was coded by two researchers individually and discussed later to achieve consistency. During this process, any new codes were added to the initial codebook with a definition. The other two researchers then reviewed the new codes and their definitions for agreement and applied the updated codebook to code the next transcript in sequence. After finishing thecoding, the authors exported codes to a spreadsheet to iteratively organize relevant codes under each research question to form subcategories and categories (see supplemental files).

## 5 RESULTS

Before identifying the violations, mods usually monitored the chat and sometimes interacted with viewers. Sometimes, they could not define whether the messages in the chat were violations. They waited for more information to evaluate the purpose and meaning of the messages. P3 (M, 23) said, *"It's difficult sometimes to ascertain things, but as long as people aren't saying, 'Oh you look awful', or things like that, I'll usually leave, and I'll try and gather more information and see what they're going to do in the chat because more often than not, I can't predict the future, at least give them the chance to talk."* For lightweight issues, several mods reported that they tended to give people chances

[3] https://www.temi.com/
[4] https://atlasti.com/cloud/

and *"watch and see"* (P12, M, 21). Once the mods confirmed the violation in the chat, the profiling process was triggered and involved evidence identification, evidence collection, and violator type formation with possible punishment.

### 5.1 Evidence Types

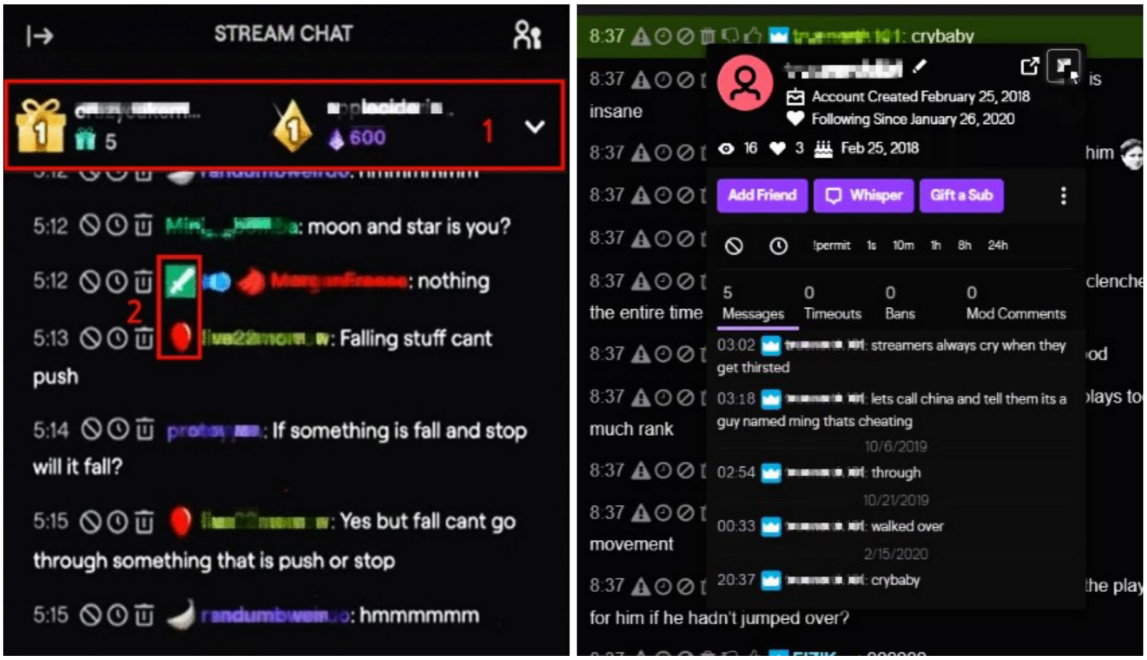
To answer the first research question, 'What kind of evidence do mods collect to profile violators?' we adopted physical evidence types [16, 17] from the criminal profiling framework and identified three types of evidence applicable to online communities. According to our observation of mods' activities, they conducted evidence collection in a very specific sequence. We followed the sequence of how mods processed information and presented this section.

*5.1.1 Action Evidence.* Action evidence refers to information that reflects the online behaviors. For example, in Figure 2, the message "fresk is bad" was considered action evidence that reflected the violator's intention and behavior to harass the streamer and was deleted. This user specifically pronounced the streamer's name and said the streamer was *"bad"*. Mods reported many different types of behaviors/acts, such as being malicious, trolling, spam, racism, and sexual perversion. These violations have been broadly discussed in prior work (e.g., [26, 31, 39]). In addition, mods also noted that disruptive behaviors such as nonsense-talking in the public chatroom broke the synchronous experience, though these messages did not break the rule. If the disruptive behaviors went far and caused trouble to other users, mods would step in and sanction these behaviors. Mods sanctioned violators differently after they turned the one-time offensive action into repeated offenses, such as P18 (F, 20): *"Sometimes they don't realize that their message is offensive, but people like that who says things impulsively. I know their intention. So I just delete it, and if they keep going, I just give a timeout."*

*5.1.2 Ownership Evidence.* Ownership evidence refers to information that reflects the identity or source of the violator. It consisted of offensive usernames and throwaways accounts, username position, badges, channel status, and account status. Though a few types of the above evidence were more or less mentioned in live streaming research (e.g., badges and throwaway accounts), we considered them necessary components to represent the holistic picture of the profiling process and explain them from the profiling and moderation perspective.

*Offensive Usernames and Throwaway Accounts.* Usernames were observable evidence that was directly and visually collected by mods. Before the violation happened, mods in most cases considered offensive usernames as heuristic indicators of the potential violation and tried to avoid their influence in the community. Offensive usernames *"indicate more that they're there to cause trouble rather than to actually participate"* (P8, F, 18). These users circumvented the rules, and the username display was too offensive to be consistent with the channel's value. For example, P12 (M, 21) shared two offensive usernames via Discord during the interview: a sexual username like *"Ice_wallo_come"* meant I swallow cum, and a sexual harassment username towards underrepresented groups like *"ray_ping_minors"* meant raping minors. Mod worked with the streamer to *"ask for them to switch over to a new account if they want to watch"* (P7, M, 18). In most cases, offensive usernames can be considered indicators of potential violators and paid special attention to these users. However, in some cases, it was context-dependent, and mods relied on other clues to figure out the purpose of users.

After the violation, an offensive username alongside a negative message (action evidence) provided additional information and enhanced mods' judgment on whether the user was an intentional violator. They noted that they would carefully check the user's account information. P2 (M, 19) expressed his logic: *"I typically immediately click a toxic name with a toxic message.*




(a) Box 1: the LeaderBoard; Box 2: badges; the icons alongside each blurred usernames are three commands: ban, timeout, delete.

(b) A moderator checks a user's message history: this user sent only 5 messages in the chat with 0 timeout, bans, and mod comments.

Fig. 1. Screenshots of the interface from a moderator's view integrated with different moderation tools.

*That makes sense."* Mods also used usernames to identify what they perceived to be throwaway accounts. P17 (M, 21) described that accounts with *"a bunch of numbers "* were *"obviously throwaway accounts"*. In order to further judge whether it was a throwaway account, P5 (F, 27) stated that she would *"go through and check their profile and see if it's like blank or anything"*. If the account history was empty, there was a high chance that the suspicious account was a throwaway account. Unlike typical throwaway accounts using letters and numbers, some accounts directly harassing others by saying something negative about a specific user or the streamer could also be considered offensive usernames.

*Username Position.* The username position on the Leaderboards [5] (a Twitch feature for the streamer to give viewers' recognition by pinning gifters' and cheerers' usernames to the top of the chat window, as shown in Figure 1a Box 1) also played a role. P3 (M, 23) explained that a big donation made the username appear on top of the chat for a certain amount of time, indicating support and contribution to the community. Mods recognized these usernames, had a higher tolerance when these users violated the rules, and were less likely to punish them, compared with users not on the Leaderboards.

*Badges.* Along with a username were the badges owned by the user (As shown in Figure 1a Box 2). Twitch offered users different types of badges [6] specific to the channels, such as cheering chat badges (users purchasing virtual currency-bits and paying bits for special animated emotes to cheer the chat and support the streamer) and subscriber badges (users paying a monthly fee to support

[5] https://help.twitch.tv/s/article/leaderboards-guide?language=en_US
[6] https://help.twitch.tv/s/article/twitch-chat-badges-guide?language=en_US

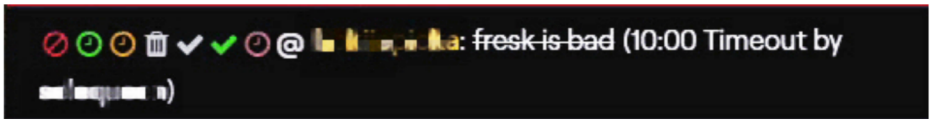



Fig. 2. A message deleted, and a user timed out by P18. The icons alongside the blurred usernames are commands without any badges, such as ban, two different timeouts, delete, etc.

the streamer and owning different badges by subscribing different lengths ). An active user usually owned various badges, either through purchasing subscriptions to the streamer or for free (e.g., VIP badge denoted by the streamer to recognize the loyal members). Usernames with badges were less suspicious than those with no badge. P12 (M, 21) described, *" The first thing that stands out to me about a user is if they have badges or not. That's like, whether you're a subscriber or if you have Twitch Prime, or if you have anything. If you have a badge, generally speaking, it's less suspicious than just a regular account with no badge."* P12 also explained that an account with no badge would make him *"curious"* and *"click"* it.

*Channel Status.* Channel status refers to the user information and activities in the channel (micro-community). Mods also reviewed channel status, specifically, following date and subscription status in the channel, by quickly clicking on the username. Subscription meant that users paid a monthly fee to support the streamer in the channel, indicating the enjoyment of the content and the contribution to the streamer. P8 (F, 18) noted, *"People typically don't throw money at the people they want to mess with and make a bad day."* Thus, the subscription was a good reflection of a user's intent. Through the observation, we asked P18 (F, 20) why she timed out a user. According to Figure 2, the user typed "fresk is bad" and got a 10-minute timeout. Fresk was the streamer's name, and the mod considered it a personal attack: *"What I do to judge is I check if the person is a sub. As you can see, he's not a subscriber as well, so I know he's not really joking."* In P18's explanation, subscription status could also be reflected by the subscriber badge alongside the username. In this example, the username had no badge and got a timeout.

Mods also checked the following date to distinguish the regular from new users. Following a channel was free and an indication of a user's interest in the stream. After following the channel, users could send messages in some follower-only chat channels. P2 (M, 19) explained, *"I would say I would click on their names, and I'll check their following age and see if they're following the person that they hosted from or see that it's a random person who just saw because of the high number of viewer count, and if it was a toxic message, and they were just following the person, then I would time them out."* In P2's sense, a short following time with toxic content suggested the user's intent to harass others. Overall, channel status indicated the loyalty and interest of the community. Mods had the mental model that users with long following time and subscription periods were valuable community members and less likely to be sanctioned.

*Account Status.* Account status refers to the user information and activities on the platform (community). Four mods stated that they would check account age that determined the user's length on this platform by clicking on the username. They consistently agreed that the account age was a good indicator of a troll account or throwaway account. We observed P3 (M, 23) checked a user's account information after a user typed "wtf" in the body painting channel and asked him why he checked and what he was looking for: *"Usually when people say something like that, I immediately think, okay, how old is the account? Do I need to ban them? But I don't believe he said anything else, so I kind of just let it go."* According to P3 and our observation, the account was created in 2019, so it was an old account, and he also checked the message history, indicating this

was the first message of the user. Thus, he *"let it go"* and gave the violator another chance. He further explained why he considered account age very important for his moderation: *"If somebody makes a throwaway account, they can just make a new one right after. They don't have to worry about, then bans technically won't even matter, but if it's an older account, more often than not, I'm less likely to ban them."* Typically, throwaway accounts were created in a short time; thus the account was new and usually untrustworthy. P7 (M, 18) also reported similar logic: *"I can check the account creation date, so I'll know if this person just made their account two hours ago. Chances are it's just like a troll account. So there's no harm if we just ban it, but that isn't to say if the person's had an account for six years, they wouldn't do something like that. So it's called context-based. I would say I'm harsher towards accounts that were recently made because I feel like you're making an account to troll, like you're going to get banned."* Similar to channel status, account status with longer age was considered more valuable to the community. Differently, channel status only reflected the activities in a specific channel, but account status reflected the account activities on the platform. The platform contained thousands of different channels. A poor channel status did not necessarily indicate a poor account status and vice versa. Mods relied on both.

*5.1.3 Sequential Evidence.* Sequential evidence refers to information that indicates the sequence of the act (e.g., chat messages with timestamps). Mods reported scrutinizing a user's message history to 1) gain context of a specific situation or a user, 2) identify the behavioral pattern, and 3) review moderation history with timestamps. The difference between action evidence and sequential evidence was that action evidence emphasized the single action reflected by the chat message, while sequential evidence emphasized the actions in the sequence.

*Chat Context (Recent Chat).* Mods often collected chat history to gain the context of a specific situation. Checking chat history facilitated their moderation actions. P19 (M, 26) stated, *"We can see their past messages. So sometimes we'll look at that and see what started the argument and like, why were they arguing with each other, why were they talking to each other?"* Similarly, P15 (M, 31) expressed that he usually *"scroll up in the chat and find out what the context is."* By comparing chat history, mods resolved the issue fairly. P13 (M, 29) described that he often went back to the message history to *"compare"* everyone's message to gain *" a little more context"* and resolved the issue. Mods also used chat history to gain an understanding of the users. P19 (M, 26) explained how he used previous messages to know users' temperament: *"I look for what they say because I never really like just anything that jumps out as toxicity or overall negativity. That's not worthy of me banning them just because of what they've previously said, but it does let me know what kind of temperament they have."* According to P19, checking previous messages was a way to understand the *"temperament"* of potential violators.

In some cases, though the users seemed to perform well, they might break the rules later. Several mods reported that if they saw single-letter expressions, which were indicators of potential spam and personal attack, they started to check the recent chat history. P15 (M, 31) said that violators used one-word messages to spell out something inappropriate, and they watched out for this type of message regularly. In rare cases, mods would like to sanction first if they lack context. P14 (F, 28) explained her moderation preference (sanctioning first, then revoking if the violator explained) and shared with us a video clip of a Whisper conversation with the violator (see Figure 3). She revoked the sanction after the violator sent a private message to explain the situation and apologized. She kindly reminded the violator to be careful and explained that the violator had only four messages in the chat history.

*Behavioral Pattern via Chat History.* Mods used chat history to identify the behavioral pattern of violators. P17 (M, 21) described his identification of a recurring troll: *"If they keep saying the*

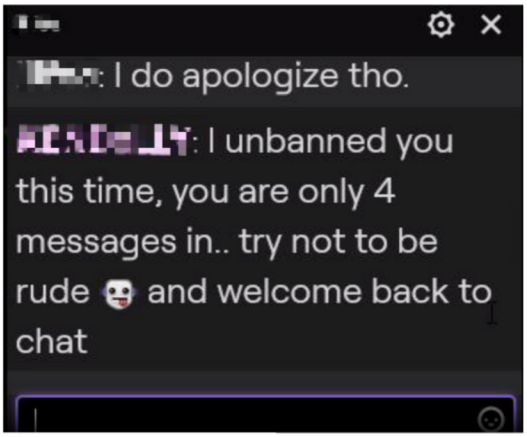



Fig. 3. A screenshot of Whisper conversation between a viewer and a mod on Twitch. After a violator apologized, the mod unbanned them with an explanation and maintained the "violator" in the community.

*same shit over and over again, like someone asks a question, right? I answered, and they ask again. That happens too many times, and then I click on his name, and I checked: 'Wait, is he like spamming the question or not?'"* In Figure 1b, P17 (M, 21) checked a user chat history after the user typed "crybaby" and explained, *"I see something in chat, and I'm like, okay, is this guy toxic like usual? Is he usually toxic? Is it like a one-time occurrence, right? So I'm looking [at] his chat log to see."* P17 found that there were only five previous messages and that this was a one-time occurrence, so he decided to let it go. The behavioral pattern not only showed what has happened but also predicted what could happen (P13, M, 29).

*Moderation History.* Moderation history included the messages being banned and timed out. Some messages were under the same moderation action, and some were even the same. These messages were repeated offenses instead of generally repeated behaviors. P12 (M, 21) described how he checked the repeated offensive messages in moderation history as references: *"I see if they've been banned before because if they're a repeat offender, I don't even think about it. They're just going to get banned again. Things like, have they sent where they banned for the exact same message before? Where they timed out for the exact same message before in a different stream? That sort of thing."* According to P12, mods checked *"the exact same"* offensive behaviors through the moderation history. The same messages guided them to sanction the violators. Interestingly, mods indicated that they also referred to evidence from different streaming channels.

Generally, among the three types of evidence, action evidence works as a trigger, ownership evidence as a start, and sequential evidence as a supplement. After seeing a negative message (action evidence), instead of commonly filtering and blocking as moderation strategies, mods first use visual cues such as username and badges to form an impression quickly (ownership evidence), then check account information to make sure whether this is a first-time violator (ownership and sequential evidence). If they lack the context, most of them would like to give users another chance and track with close attention (sequential evidence), waiting for more evidence to understand the context and intent.

### 5.2 Evidence Collection

To answer the second research question: how mods collect these types of evidence, we found that mods collected these types of evidence in five different ways, including documenting, co-experiencing with viewers as inference, collaborating with moderation teams (across channels),

gaining knowledge from users by staying in the community for a long time, and relying on moderation tools.

*5.2.1 Documenting.* A few mods stated that they shared a spreadsheet containing violators' information with notes in the moderation team. They also did cross-channel documenting, which meant several channels individually documented the violators and shared with others. Documenting was a way to mainly collect ownership evidence. P15 (M, 31) described how the information collected on a Google document helped him gain context of the violation: *"We have a shared Google document. It has a list of not finding the word, but people that have caused problems in the past for timeout or ban or whatever. If I lack context in a situation in that community, then I can go to that spreadsheet. I can search for that person's name, and I can see if they've been a problem in the past or if this is their first infraction."* Furthermore, P15 stated that they also shared the document across different moderation teams so that other channels could pay close attention to these violators: *"We have a sheet that is just for known troublemakers, so moderation teams from other streams will see. These are the people that we had issues with. Here are their usernames so that you can be aware, and then somebody comes in, and they start saying something that they might seem to be innocent at first, but we know based on the information from another moderation team that this is someone who has been a problem in the past, so we can watch out for them if they start to go down a path of being a troll or whatever."* According to P15, mods applied external platforms that were not initially designed for moderation to collaboratively moderate, either within the channel or across different channels.

*5.2.2 Co-experiencing with Viewers as Inference.* Some mods reported that they were viewers and watched other channels that streamed similar content to their moderated channel. Similar streaming content attracted similar types of viewers. Thus, they knew the background and actions of violators in other channels. When these violators came into their channels, they recognized them. For example, P7 (M, 18) stated, *"There're also people from other streams that come in, I know from their stream, their respective place."* P13 (M, 29) described how he recognized viewers from other streams through ownership evidence: *"Some people have some pretty weird names, right? You can kind of see. When you see the kind of stuff that you don't really think is right, you kind of subconsciously remember it a little more."*

*5.2.3 Collaborating with Moderation Teams (Across Channels).* Many mods also reported that they collaborated with other mods in either the team of the channel or teams across channels in mainly three ways (asking other mods' opinions within the channel, cross-channel log check, and multiple channel moderation ). The nuanced difference between cross-channel log check and cross-channel documenting was that log check included all chat history while documenting only included violation behaviors. The difference between co-experiencing with viewers as inference and multiple channel moderation was that mods were viewers in other channels in the former situation and were moderators in other channels in the latter situation.

Some mods asked other mods' opinions, like sending a *"screenshot of the message or log"* (P18, F, 20) to others when they lacked background information of a particular viewer, in line with prior work that mods had group discussion during the moderation process [66]. A few mods stated that they had a collaboration with other streamers and could conduct cross-channel log check through third-party platforms. P14 (F, 28) described how her team applied a third-party platform called Overrustlelog [7], a public chat log website for Twitch channels, to collaborate with other streams: *"I know all of our mods, we do this like, for example, a lot of people don't like XXX. She's another streamer, and we've had to ban a few of our people that went over to her chat to be toxic. We know this because we saw in her Overrustlelog, so it wasn't just hearsay ... We've had to cross ban people that*

[7] https://overrustlelogs.net/

*got from our community had gone over to her chat to be douchebags, and so we'll ban them in our chat."* According to P14, mods sometimes moderate not only violators within their channels but also users who were considered violators in other channels though they did not break the rule in their channels.

A few mods also noted that they moderated across different channels sharing similar viewers, and the viewers' behaviors in other channels could be indicators of their decisions of the current channel. P10 (M, 18) shared his experience moderating two streams with the same content: *"When both streamers are live, or when they were streaming together, like playing this together, you would have viewers in one chat that are toxic in one chat that would obviously be toxic in the other. Since I'm a moderator for both, it's kinda clear. I remember viewers from one chat that break the rules a lot."* Some violators kept the same username across different channels; mods easily remembered their names. P8 (F, 18) stated: *"A lot of the time people that are there to cause problems, they don't change accounts. They just keep the same name, so what you find in maybe one person's stream, you might ban them, and then you might see their name a few hours later, and you'll go, I remember that name."* Both P10 and P8 in common described they remembered violators' names in a short time, either at the same time or *"a few hours later"*. Team collaboration mainly relied on usernames as references and violation history to gain context. Mods also expressed the challenge of identifying violators if they completely changed their usernames across different channels.

*5.2.4 Gaining Knowledge from Users by Staying in the Community for a Long Time.* Some mods stated that they had been in similar communities for a long time and recognized viewers through frequent seeing. P15 (M, 31) said, *"I've been in this channel for seven years at this point. You spent time in channels over time, you learn the regulars, you get to know them, and you recognize them."* Similarly, P6 (F, 34) said some users actively appeared in Twitch chat and Discord channel to interact with others, and mods *"kind of know how long they've been around"*. Combining the frequent seeing of usernames with other evidence helped mods figure out the intent of users. P16 (M, 24) explained, *"So they've subscribed to XXX, I think that's a five or six-year badge, so that's a lot of money to give to XXX. I've seen their names a lot. I can tell that they like XXX, So if that person types the same message, I fucking hate you, and I'm going to probably understand it as 'oh he's jokingly hating the person.'"* According to P16, mods combined account status, badges, and frequently seeing users to interpret users' behaviors, finding out that this user was *"jokingly hating the person"*.

*5.2.5 Relying on Moderation Tools.* Most mods applied various bots in addition to the AutoMod offered by Twitch to facilitate the moderation process. Many mods applied third-party tools, such as Better Twitch TV and FrankerFaceZ, to customize moderation action, similar to prior work [5]. As shown in Figure 2, there was a list of customized buttons in front of the username. At the same time, tools allowed mods to collect various types of evidence such as account age, channel status, and message history in the channel. For example, P5( F, 27) sometimes *"go through and check their profile"* to determine throwaway accounts with the assistance of moderation tools. They mainly collected ownership evidence and sequential evidence, Figure 1b showed a typical interface of Twitch AutoMod. This account was created in 2018, indicating that it was an old account. It followed this channel in 2020, several months ago. Bans and timeouts were "0", indicating it might be a good user. Moderation tools provided the necessary information to help mods form the first impression on users quickly.

### 5.3 Types of Violators

To answer the third research question, 'What are the types of violators that mods perceive?' we identified five types of violators. Moreover, *"racist"* and *"sexist"* were commonly mentioned by mods with a consistent attitude towards sanctions. They are easily recognized via action evidence

and sequential evidence, with the assistance of ownership evidence. Mods would ban them without further consideration. We present the other five types of violators reflecting mods' complex attitude and decision-making process.

*5.3.1 Violators Performing Malicious Mischief.* Criminal mischief, also called malicious mischief, refers to behaviors intentionally damaging another person's property in criminal justice. Several mods reported a type of violator who randomly came into a channel to cause trouble and intended to disrupt the community. For example, P15 (M, 31) said, *"You have people who come in ,and they just want to be malicious. They come in specifically to be disruptive. They come in specifically to cause an issue, to force the mod team to do something."* Similarly, P6 (F, 34) expressed that this type of violator wanted to see the anger from the streamer: *"I think they just want to get a rise out of the streamer. They want the streamer to kind of fightback there."* This type of violator took advantage of the anonymity and pseudonymity of the Internet and obtained excitement from the mischief. P13 (M, 29) said, *"Maybe they just appreciate the anonymity or that, and they're just like, 'hey, we can haha, we can get a rise out of people if we do this.'"*

*5.3.2 Attention and Reaction Seekers.* Many mods mentioned that a type of violator was the attention and reaction seeker. Unlike violators performing malicious mischief, these attention and reaction seekers did not initially try to cause trouble. They competed for recognition mainly through sarcasm and troll and for popularity through self-promotion.

*Attention Seekers Needing Recognition.* Some violators wanted to *"get recognition from a streamer"* (P6, F, 34). *"They're trying to get people to notice them, to validate them and their actions, so it's not always because they disliked the stream or they disliked the viewers. It's because they need to be seen and recognized, and they have the need to be validated,"* said P15 (M, 31). These violators broke the rules because they had the desire to be recognized by others. Once their needs were recognized and fully fulfilled, violators might *"turn to normal people"* ( P19, M, 26). However, the overwhelming messages made the streamer not recognize them. P4 (M, 18) explained that *"everybody wants attention"* and said, *"Because they like watching the streamer, so they want attention from the streamer, reading the question, answering it or saying hello to them. It's a personal connection through the screen."* In P4's sense, attention for recognition was considered a strong personal connection with the streamer. Massive viewers wanted to be recognized by the streamer, thus forming completion. In order to stand out, some violators attempted to be sarcastic or make trolls. P3 (M, 23) suggested that some sarcastic jokes were in the *"grey area"*. Thus, mods needed to put it into the context of the conversation to interpret its meaning. For example, P19 (M, 26) told us that they could usually differentiate whether it was a sarcastic or toxic comment: *"We can usually tell because there'd be other types of comments in there. There'll be conversational comments with other chatters. There'll be other statements about stream ... Those would normally be considered a toxic comment, then put into context of what they love, what other things they said, and you realize it's potentially not toxic. It's potentially just sarcasm."* According to P19, mods mentally categorized comments into different types and applied the chat history as a context to interpret the underlying meaning of the messages. Mods relied mostly on sequential evidence to make the judgment. In extreme cases, some violators experienced mental health issues in offline life and started the *"psychological cry for help"* (P15, M, 31) online. P16 (M, 24) explained that he believed the violators were not negative offline, and most violators did not have an outlet to let all anger and depression out in real life, so they came to online communities.

*Attention Seekers Seeking Popularity.* Another type of attention seeker was violators who wanted to gain popularity by promoting themselves in other streams. P4 (M, 18) described that some streamers (competitors) in small channels went to the big channels to post advertisements and

*"make as much noise as they can"*. Gaining popularity was the main reason, and *"attention, popularity, intention kind of go hand in hand "*. P4 added, *"They can make a disturbance and say, you know, go follow me on this, on their social media sites, or they'll shout themselves out in front of thousands of people in chat. That's obviously not acceptable."* Similarly, P6 (F, 34) noted, *"You have the kind of attention seekers who will hop in and be like, 'hey, look at me. I'm a streamer to those.' We don't like those. I don't want other people advertising, so we get rid of them."* P4 and P6 indicated this type of attention seeker were not *"acceptable"* and would like to *"get rid of them"*.

*5.3.3 Immature Juvenile.* Four mods mentioned juvenile as a type of violator and usually treated it differently. P8 (F, 18) explained that she moderated in years and could *"pick up on the pattern"* to identify juvenile violators through messages and tones they used: *"He tends to talk in caps with very bad grammar and you can kind of look at that and go, 'Oh, that is more than likely a little kid rather than a problem.' He also thinks like the most random things are funny ... You can tell that they think it's really funny, and that tends to be more of childish humor. It doesn't mean everyone with that humor is a child, but it leans more towards being a child."* According to P8, juveniles preferred using capital letters with bad grammar. The language pattern also indicated that juveniles and adults had different senses of humor. P15 (M, 31) supported P8's explanation: *"They'll use the letter U instead of the word YOU, they'll use the letters UR instead of YOUR, and they'll do a lot of things like that to make it abbreviated, to use what people called 'tech speak.'"*

After mods identified the pattern of juveniles, they preferred communication to sanctioning. P8 said, *"We try to talk to them more than actually take action against them because we're trying to help them understand why what they're doing isn't proper."* P15 further shared an example: *"He went straight to saying very inappropriate things about the streamer and [the streamer] talked to him, asked him what was going on, and I ended up stepping in and talking to him, asked him if he needed to talk, ended up talking to the kid for a couple of hours that night and come to find out his parents were going through a divorce, and his dad had abused his mother that day and then left the house, so he was upset. He didn't know how to properly vent his feelings, and his way was to go onto Twitch and try and be a troll. So ended up talking to him for a few hours, and then he became an active member of the community for a couple of years after that."* According to P15, consistently, mods reported that some violators experienced mental health issues and used online communities to vent emotions they suffered from offline life. Mods tended to have a strong tolerance for juveniles' violations and would like to help. In this sample, communication helped the juvenile and transferred the violator to an active community member.

*5.3.4 Repeated Offenders with Contributive Participation.* Mods stated that some violators were toxic regulars but also active community members. These violators kept breaking the rules, accepting punishments, still staying in and contributing to the community, and breaking the rules later. These violators were *"stubborn"* and *"unable to adapt or change"* (P12, M, 21) but valuable community members. P16 (M, 24) said, *"The part that makes me like them is that they do actually interact with the chat room. It's like they talk to each other. They talk to the streamer. Occasionally with frequency, they will break the rules. It's like a very nice criminal that you consistently arrest, but they're always respectful to you. They're respectful to the content of the streamer. They're respectful to the streamer. They're respectful to everybody else in the chat room, but they just have this habit of getting in trouble."* According to P16, some violators having the *"habit of getting in trouble"* are active community members and *"respectful"* to the community. These violators are considered *"nice criminals"* because they accepted the mistakes they made and the sanctions that they were given with no intention to leave the community. Some mods had mixed feelings and concerns about the punishment for this type of violator. Generally, they sanctioned them differently, considering their contributions. P9 (M, 19) shared his experience moderating "active" but also "toxic" viewers: *"So it's very difficult to*

*decide how we're going to deal with them because they're still a very active part of the community. They're contributing a lot to the community. It's just like, occasionally they make mistakes that are against the rules, but we punish them differently because they're adding a lot to the community and they're like helping. So it kind of gets hard to figure out what sort of punishment we're going to give them.*" According to P9, mods sanctioned the repeated offender and repeated offenders with active participation differently and experienced difficulty in deciding the sanction level to this type of violator.

*5.3.5 Aggressive and Hostile Attackers.* Another type of violator was viewers who were aggressive and hostile. This type of violator could be easily triggered to start harassing or attacking others. P12 (M, 21) said, *"We have viewers who come in, and they do stuff they're not supposed to do, and then they get timed out for it, not necessarily banned, but then they get aggressive. So they'll either in my whispers, 'why'd you time me out? You're a piece of shit. Like kill yourself'... or after the 10 minutes they'll come back and chat, 'Wow, your mods are absolutely trash, blah, blah, blah, like fuck you.'"* In this case, the violator was not satisfied with the moderation and started the aggressive behaviors through either Whisper or the public chat to attack the mod. Some violators who got banned in Twitch communities targeted other relevant communities to continue the attack. P14 (F, 28) shared an example of a violator posting on the subreddit of the Twitch channel to accuse that mods abused the power of banning people, and then the Twitch mods and the violator started the argument on Reddit. In other cases, if mods understood violators' personalities and knew their intent, they might allow it. P6 (F, 34) said, *"I know we've had one person that's been a regular, and she'll often do the backhanded threat of 'I will cut you'. We know she's not going to, but it's more of that feisty spirit more than anything. So it's like, yeah, we know she's not really going to attack this person."* According to P6, though this violator threatened other viewers, the moderator knew this violator and considered the violation behavior not serious enough to warrant punishment, whereas someone else who said the same thing might have been subject to a different type of sanction.

## 6 DISCUSSION

We use criminal profiling as a lens to guide us to understand the mental model of mods who have a non-expert profiling background when they deal with potential violators. Mods work as both evidence collectors and profilers in the moderation process. We find that mods mainly collect three types of evidence in five different ways. The five methods of collecting evidence mainly rely on individual experience and collaborative work with limited technical support from the platform, mostly collecting ownership evidence and sequential evidence. After the evidence collection, mods unconsciously fit violators into mainly five types and apply different moderation strategies.

We clarify that the mental model in this work consists of two parts: the first is about collecting and using different evidence; the second is about the types of violators requiring different moderation strategies. The pattern of evidence types and collection might generalize to other online communities that aim to thrive via extensive effort in the moderation process. The different affordances of platforms might cause the process to be a little different. For example, on asynchronous platforms such as Twitter and Reddit, users' activities such as posts and replies are saved under a user's profile, making the evidence collection process comparably easy. Content removal and banning users are easy and sometimes can effectively decrease toxicity from existent users but force other users to migrate to other platforms [12]. The types of violators identified in this work show the complexity of users' behaviors . These types provide community administrators an alternative to consider punishment if they aim to maintain community members. Meanwhile, commercial moderation teams who work for social media and news sites might integrate the mental model into the moderation process and use it to restrain severe sanctions for first-time offenders.

### 6.1 Platform Design and Affordance Make Profiling Go Beyond the User's Profile

A user's profile often contains registration information and account activities. Prior work has explored how users on social media sites curate self-presentation to maintain social relationships with other users through different profile elements, such as a profile image [83, 84], the about me and interest [37, 52], and the location field [38, 77]. Account activities under a user's profile provide cues to understand the user. For example, peers in the open-source community form impressions about other users' expertise based on the history of activity across projects and the successful collaboration with key high-status projects [54].

In live streaming communities, we find that many mods frequently mention they check the account status and channel status because of the limited information on a user's profile. The interface of a user's homepage is initially designed for those who will be streamers. We speculate that the design discourages viewers from filling in the relevant information. In addition, different from posts and feeds on social media sites like Facebook and Twitter, activities such as message histories and replies are not stored under the user's profile from the user's end because of the synchronicity and ephemerality of the "live" affordance. In other words, after the streamer closes the stream or the user leaves the channel, the user cannot store or see the message history in the channel anymore. Once the users leave and come back, the message history is erased and displayed from the time point the user gets in. In addition, mods have to apply tools to log chat histories of a user only in the specific micro-community/channel. Mods in the current micro-community cannot see the message history and violation in other ones. Mods also do not have access to log data of other micro-communities. The limited information on a user's profile and the challenges of acquiring other information compel mods to seek other methods to collect evidence beyond a user's profile and across various micro-communities. Thus, understanding evidence collection is essential to figure out the profiling.

### 6.2 Profiling as Part of the Moderation for Community Growth

Different from the goal of criminal profiling for crime capture [25], in online communities that include thousands of micro-communities, the goal of violator profiling is to avoid punishing users, even help users in some cases, to grow the micro-communities. In our observation, mods rarely directly ban users only based on the content. Even when they do so, they can easily revoke after users express remorse and apology. On Twitch, mods frequently play roles to facilitate the community, such as facilitator, mediator, and adult in the room [62]. Overall, mods are willing to go the extra mile to retain community members.

*6.2.1 Fairness and Justice.* In criminal justice systems, retributive justice suggests sanctioning violators with proportional punishment for their violations [8] and is predominantly applied on commercial platforms. Most volunteer mods in user-governed micro-communities show a preference for restorative justice, involving the repair of justice by bringing stakeholders such as violators, victims, and mediators together to acknowledge and remediate harm [78]. Prior work shows that retributive justice is not the most effective measure to promote reconciliation, and restorative justice can potentially complement it to initiate and boost reconciliation [2, 18]. Mods in live streaming communities often work as facilitators to mediate the conflict in the chat, such as asking users to change or stop a particular behavior and helping users with trouble in offline life. Profiling as part of the moderation process in live streaming communities shows an example of the application of restorative justice to users, supplementing recent work appealing a restorative justice to support targets of harassment online [60].

The complex behaviors for each type of violator indicate the same standards of punishment to these violators are considered unfair and unjust. The one-fit-all approach will fail and drift away

from these potentially valuable users [60]. After profiling, mods identify the types of violators who need help or unintentionally break the rules. Mods choose to communicate with and take care of them instead of outright punishing them. The caretaking and restorative approach make these one-time or one-day violators become loyal community members later. Accordingly, sanction after profiling could potentially increase the perceived fairness and justice in these spaces.

*6.2.2 Bad Act and Bad Actors.* Profiling allows mods to ascertain the user as a bad actor not only based on the bad act at the scene. Our work supplements prior work arguing that moderation should consider the context [9] and reveals some more sophisticated scenarios. Mods consider not only the content and context but also the violator's intent and characteristics, sometimes their experience, into moderation. Many mods describe they apply the other channel violation as a reference of moderation action in the current channel; in rare cases, mods rely on what happened in other channels as a way to understand a user's personality, not a reference of sanction.

Automated moderation systems heavily rely on the content and consider the bad act as a violation and sanction bad actors. Our results show that though mods recognize the "bad" actors, it is difficult for them to assign the punishment in some situations. For example, mods express that they weigh the violators' contribution and tend to have more tolerance to repeated offenders with contributive participation. Notably, many mods consistently express emotional and social support to immature juveniles and would like to talk with and educate them. They also sanction similar behaviors differently for other types of violators. For example, knowing the aggressive and hostile attackers' personalities and intent is critical for mods to decide approval or ban; attention seekers seeking popularity are directly banned, but those needing recognition depend. Profiling in the moderation process attempts to decrease the bias and discrimination created by the automated moderation system [3, 32] and allows mods to distinguish the bad actors from the "bad" act.

### 6.3 Implications and Recommendations

We propose designs to facilitate collaborative and individual violator profiling and to integrate violator profiling into the moderation system that combined automated and human processes.

*6.3.1 Build a Mechanism for Collaborative Profiling.* Mods are collaboratively getting rid of violators, collecting violators' information across different micro-communities via external tools or platforms not essentially designed for profiling. According to the official website of OverRustleLogs, it was shut down in May 2020 at the request of the Twitch Legal team because of privacy concerns. However, the internal tools only work for a specific channel. We suggest the platform develop a mechanism that allows all mods at the micro-community level to list violators and to share the information with other mods at the community level. The pseudonymity of Twitch helps reveal more information of violation while keeping violators' real identities safe.

Cross-channel collaborative moderation also indicates the possibility of cross-platform moderation, Tech giants (Facebook, Microsoft, Twitter, and YouTube) together established the *"Global Internet Forum to Counter Terrorism"* in 2017 to coordinate content removal about "violent terrorist imagery and propaganda" [35]. However, there is little or no collaboration about dealing with daily online harassment. We propose a mediated system that allows different online communities to document and share violators' information to the commercial moderation teams or volunteer moderation teams. Though commercial moderation teams work behind the scene and are managed by the platform [58], which can protect the violators' information privacy while allowing mods to deal with the violation, we don't know how to keep the boundary between privacy and profiling in the volunteer moderation teams, requiring further investigation.

*6.3.2 Facilitate Individual Mod to profile Violators.* Some recommendations to facilitate individual profiling should be highlighted to supplement collaborative profiling. First, we suggest a mechanism allowing mods to label and tag violators manually. Recent work has developed prototypes to use algorithms to analyze the message history to automatically label users [41] and summarize messages as key points [82]. Our findings reveal the complexity of violators' characteristics. Thus, we suggest integrating a mechanism (either developed by Twitch or third parties) containing a database with pre-defined personality traits and violator types in criminology and psychology. These labels might help mods scrutinize factors that are not achievable by algorithms and allow them manually tag violators.

Second, we suggest a feature allowing mods to trace username change. Mods explain that sometimes they can remember the usernames or recognize the users, but the vague memory and change of usernames increase recognition difficulty. The process is primarily supported by social, not computational practice, making the recognition very random. The current AutoMod allows mods to log the message history of users. The pseudonymity of online communities encourages self-disclosure and free speech [76] but also increases the profiling challenge. We suggest developing a feature in the moderation tool that can trace the username change history across different micro-communities. These designs align with the current Twitch moderation mechanism, which is only visible to the streamer and mods to facilitate the moderation process.

*6.3.3 Resource for Training and Educating Mods.* Prior work shows professional profilers can produce a more accurate prediction of an unknown offender, comparing to other groups [49]. We find that mods, as non-experts in profiling, own much power to sanction violators, and the process sometimes is pretty subjective, varying from person to person. Platforms might offer resources for training and educating mods to avoid false profiling, such as making online video tutorials to explain the importance of profiling and integrating the components into the moderation guideline to show mods how to profile step-by-step.

### 6.4 Limitation and Future Work

This work suffers several limitations. First, the data collection is from a single platform — Twitch, which is different from other asynchronous communities. Future work should do cross-platform research to validate the findings. Second, our participants were mainly in Europe and North America, but live streaming service is also booming in Asia [53]. Future work can apply our findings in a cross-cultural context. Third, our participants are mods who are willing to share the video and content; thus, we may have recruited mods who are more inclined toward restorative justice. We do not know the justice preference of the mods who are unwilling to share content. Moreover, whether the recording task affects mods behaviors needs further investigation. Fourth, though we show profiling violators as a phenomenon in live streaming communities, we can not answer questions like how frequently mods use profiling in the real-time context. Future research can apply quantitative methods with log data to explore this question. Additionally, streamers' characteristics (e.g., gender, age, preference) and the channel characteristics (e.g., content categories, community size, clarity of rules) might also have significant effects on how and when mods will choose to use profiling during the moderation process. Future work can incorporate these characteristics into algorithmic models to investigate their relationships. Last, we don't know if the profiles that moderators create are accurate representations of the violators, as we only focused on moderators' thought and behavioral processes in constructing these profiles. Future research may want to see if these profiles are accurate assessments.

## 7 CONCLUSION

In this work, we aimed to understand how volunteer mods on Twitch create profiles of violators before they decide on what action they will take with the violator. We found that profiling improved mods' understanding of violators, and they engaged in complex practices of evidence collection and documentation to create these profiles. These practices happened not just within one community but across different Twitch communities as well as on different platforms.

Generally, instead of sanctioning violators, mods preferred to go the extra mile to integrate the violators into the communities. Though they had to sanction some violators, the profiling led to different sanction decisions. We also found that mods across different micro-communities collaboratively worked on violator profiling because of the limited information in the user's profile and limited technical support from the platform.

## ACKNOWLEDGMENTS

This research was funded by by National Science Foundation (Award No. 1928627). Thanks to *ALL* the anonymous reviewers and Cody Buntain for their insightful feedbacks. Thanks to the research assistants in SocialXLab at NJIT (Andrew Suarez, Aaron Samuel, Abdelmalek Benaissa, Jessy Martinez) for data collection.

## REFERENCES

[1] Anti-Defamation League. 2020. *Online Hate and Harassment: The American Experience 2020.* Technical Report.

[2] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When online harassment is perceived as justified. In *12th International AAAI Conference on Web and Social Media, ICWSM 2018*. 22–31. www.aaai.org

[3] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from HeartMob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 24. https://doi.org/10.1145/3134659

[4] Lindsay Blackwell, Mark Handel, Sarah T. Roberts, Amy Bruckman, and Kimberly Voll. 2018. Understanding "bad actors" online. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2018-April. 7. https://doi.org/10.1145/3170427.3170610

[5] Jie Cai and Donghee Yvette Wohn. 2019. Categorizing Live Streaming Moderation Tools. *International Journal of Interactive Communication Systems and Technologies* 9, 2 (7 2019), 36–50. https://doi.org/10.4018/ijicst.2019070103

[6] Jie Cai and Donghee Yvete Wohn. 2019. What are efective strategies of handling harassment on twitch? Users' perspectives. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. 166–170. https://doi.org/10.1145/3311957.3359478

[7] Jie Cai, Donghee Yvette Wohn, and Mashael Almoqbel. 2021. Moderation Visibility: Mapping the Strategies of Volunteer Moderators in Live Streaming Micro Communities. In *ACM International Conference on Interactive Media Experiences*. ACM, New York, NY, USA, 61–72. https://doi.org/10.1145/3452918.3458796

[8] Kevin M. Carlsmith and John M. Darley. 2008. Psychological Aspects of Retributive Justice. *Advances in Experimental Social Psychology* 40 (2008), 193–236. https://doi.org/10.1016/S0065-2601(07)00004-4

[9] Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms matter: Contrasting social support around behavior change in online weight loss communities. In *Conference on Human Factors in Computing Systems - Proceedings*. 1–14. https://doi.org/10.1145/3173574.3174240

[10] Stevie Chancellor, Jessica Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. Thyghgapp: Instagram content moderation and lexical variation in Pro-Eating disorder communities. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, Vol. 27. 1201–1213. https://doi.org/10.1145/2818048.2819963

[11] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. CrossMod: A cross-community learning-based system to assist reddit moderators. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. 30. https://doi.org/10.1145/3359276

[12] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22. https://doi.org/10.1145/3134666

[13] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2. CSCW, 25. https:

//doi.org/10.1145/3274301
[14] Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of blocked community members: redemption, recidivism and departure. In *Proceedings of the 2019 World Wide Web Conference*. 12. https://doi.org/10.1145/3308558.3313638
[15] Ed H Chi. 2004. Transient User Profiling. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 2–5. http://www.geekbiker.com
[16] W. Jerry Chisum and Joseph M. Rynearson. 1997. *Evidence and Crime Scene Reconstruction* (5th ed.). National Crime Scene Investigation and Training, Redding, CA.
[17] W. Jerry Chisum and Brent E. Turvey. 2011. Methods of Crime Reconstruction. In *Crime Reconstruction* (2 ed.). Academic Press, Chapter 8, 179–209. https://doi.org/10.1016/C2010-0-67906-5
[18] Janine Natalya Clark. 2008. The three Rs: retributive justice, restorative justice, and reconciliation. *Contemporary Justice Review* 11, 4 (2008), 331–350. https://doi.org/10.1080/10282580802482603
[19] Maxime Clément and Matthieu J. Guitton. 2015. Interacting with bots online: Users' reactions to actions of automated programs in Wikipedia. *Computers in Human Behavior* 50, 1 (2015), 66–75. https://doi.org/10.1016/j.chb.2015.03.078
[20] Link Daniel, Hellingrath Bernd, and De Groeve Tom. 2013. Twitter integration and content moderation in GDACSmobile. In *Proceedings of the 10th International ISCRAM Conference*. 67–71. http://publications.jrc.ec.europa.eu/repository/handle/111111111/32413
[21] Paul B. de Laat. 2016. Profiling vandalism in Wikipedia: A Schauerian approach to justification. *Ethics and Information Technology* 18, 2 (6 2016), 131–148. https://doi.org/10.1007/s10676-016-9399-8
[22] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1–13. https://doi.org/10.1145/3290605.3300372
[23] Craig Dowden, Craig Bennell, and Sarah Bloomfield. 2007. Advances in Offender Profiling: A Systematic Review of the Profiling Literature Published Over the Past Three Decades. *Journal of Police and Criminal Psychology* 22, 1 (2007), 44–56. https://doi.org/10.1007/s11896-007-9000-9
[24] Janna Lynn Dupree, Richard Devries, Daniel M. Berry, and Edward Lank. 2016. Privacy personas: Clustering users via attitudes and behaviors toward security practices. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 5228–5239. https://doi.org/10.1145/2858036.2858214
[25] Steven A. Egger. 1999. Psychological profiling: Past, Present, and Future. *Journal of Contemporary Criminal Justice* 15, 3 (1999), 242–261.
[26] Casey Fiesler, Jialun "Aaron" Jiang, Joshua McCann, Kyle Frye, and Jed R Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. 72–81. https://doi.org/10.1016/j.tvjl.2007.05.023
[27] Andrew T Fiore, Lindsay S Taylor, G A Mendelsohn, and Marti Hearst. 2008. Assessing attractiveness in online dating profiles. In *Conference on Human Factors in Computing Systems - Proceedings*. 797–806. https://doi.org/10.1145/1357054.1357181
[28] Bryanna Fox and David P Farrington. 2018. What have we learned from offender profiling? A systematic review and meta-analysis of 40 years of research. *Psychological Bulletin* 144, 12 (2018), 1247–1274. https://doi.org/10.1037/bul0000170
[29] Vernon J. Geberth. 1981. Psychological Profiling. *Law and Order* 29, 9 (1981).
[30] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media and Society* 20, 12 (2018), 4492–4511. https://doi.org/10.1177/1461444818776611
[31] Debbie Ging and James O'Higgins Norman. 2016. Cyberbullying, conflict management or just messing? Teenage girls' understandings and experiences of gender, friendship, and conflict on Facebook in an Irish second-level school. *Feminist Media Studies* 16, 5 (2016), 805–821. https://doi.org/10.1080/14680777.2015.1137959
[32] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society* 7, 1 (2020), 1–15. https://doi.org/10.1177/2053951719897945
[33] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law and Technology* 17, 1 (2015), 68. https://digitalcommons.law.yale.edu/yjolt/vol17/iss1/2
[34] Yulong Gu, Zhuoye Ding, Shuaiqiang Wang, and Dawei Yin. 2020. Hierarchical user profiling for e-commerce recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 223–231. https://doi.org/10.1145/3336191.3371827
[35] Chloe Hadavas. 2020. The Future of Free Speech Online May Depend on This Database. *Slate* (8 2020). https://slate.com/technology/2020/08/gifct-content-moderation-free-speech-online.html
[36] William A Hamilton, Oliver Garretson, and Andruid Kerne. 2014. Streaming on twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing*

*Systems*. 1315–1324. https://doi.org/10.1145/2556288.2557048
[37] Jeffrey T Hancock, Catalina Toma, and Nicole Ellison. 2007. The truth about lying in online dating profiles. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 449–452. https://doi.org/10.1145/1240624.1240697
[38] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. 2011. Tweets from justin bieber's heart: The dynamics of the "location" field in user profiles. In *Proceedings of ACM Conference on Human Factors in Computing Systems*. 237–246. https://doi.org/10.1145/1978942.1978976
[39] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for safety online: Managing "trolling" in a feminist forum. *Information Society* 18, 5 (2002), 371–384. https://doi.org/10.1080/01972240290108186
[40] Scotia J. Hicks and Bruce D. Sales. 2007. *Criminal profiling: Developing an effective science and practice*. American Psychological Association. 293 pages. https://doi.org/10.1037/11428-000
[41] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized Social Signals: Computationally-Derived Social Signals from Account Histories. In *Proceedings of the Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3313831.3376383
[42] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did you suspect the post would be removed?": Understanding user reactions to content removals on reddit. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. CSCW, 33. https://doi.org/10.1145/3359294
[43] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction* 26, 5 (2019), 35. https://doi.org/10.1145/3338243
[44] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter?: User behavior after content removal explanations on reddit. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. 27. https://doi.org/10.1145/3359252
[45] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Transactions on Computer-Human Interaction* 25, 2 (2018), 1–33. https://doi.org/10.1145/3185593
[46] Jialun Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation challenges in voice-based online communities on discord. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. 23. https://doi.org/10.1145/3359157
[47] Charles Kiene, Jialun Jiang, and Benjamin Mako Hill. 2019. Technological frames and user innovation: Exploring technological change in community moderation teams. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. CSCW, 23. https://doi.org/10.1145/3359146
[48] Charles Kiene, Kenny Shores, Eshwar Chandrasekharan, Shagun Jhaver, Jialun Jiang, Brianna Dym, Joseph Seering, Sarah Gilbert, Kat Lo, Donghee Yvete Wohn, and Bryan Dosono. 2019. Volunteer work: Mapping the future of moderation research. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. NY, USA, 492–497. https://doi.org/10.1145/3311957.3359443
[49] Richard N Kocsis. 2003. Criminal Psychological Profiling: Validities and Abilities. *International Journal of Offender Therapy and Comparative Criminology* 47, 2 (2003), 126–144. https://doi.org/10.1177/0306624X03251092
[50] Richard N. Kocsis, Harvey J. Irwin, Andrew F. Hayes, and Ronald Nunn. 2000. Expertise in psychological profiling: A comparative assessment. *Journal of Interpersonal Violence* 15, 3 (2000), 311–331. https://doi.org/10.1177/088626000015003006
[51] Cliff Lampe and Paul Resnick. 2004. Slash(dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*. 543–550. https://doi.org/10.1145/985692.985761
[52] Cliff A.C. Lampe, Nicole Ellison, and Charles Steinfield. 2007. A familiar face(book): Profile elements as signals in an online social network. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 435–444. https://doi.org/10.1145/1240624.1240695
[53] Zhicong Lu, Haijun Xia, Seongkook Heo, and Daniel Wigdor. 2018. You watch, you give, and you engage: A study of live streaming practices in China. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. 1–13. https://doi.org/10.1145/3173574.3174040
[54] Jennifer Marlow, Laura Dabbish, and Jim Herbsleb. 2013. Impression formation in online peer production: Activity traces and personal profiles in github. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. 117–128. https://doi.org/10.1145/2441776.2441792
[55] Aiden McGillicuddy, Jean Grégoire Bernard, and Jocelyn Cranefield. 2016. Controlling bad behavior in online communities: An examination of moderation work. In *2016 International Conference on Information Systems, ICIS 2016*. 11. https://slashdot.org/moderation.shtml
[56] Stuart E Middleton, Nigel R Shadbolt, and David C De Roure. 2004. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems* 22, 1 (2004), 54–88. https://doi.org/10.1145/963770.963773

[57] NFSTC. 2013. *A Simplified Guide to Crime Scene Investigation.* Technical Report. National Forensic Science Technology Center. http://www.forensicsciencesimplified.org/csi/how.html
[58] Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. In *The Intersectional Internet: Race, Sex, Class and Culture Online*, Safiya Umoja Noble and Brendesha Tynes (Eds.). NY: Peter Lang, New York, Chapter Commercial, 147–160. https://doi.org/10.1007/s13398-014-0173-7.2
[59] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, Shammur A Chowdhury, and Bernard J Jansen. 2020. A Literature Review of Quantitative Persona Creation. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3313831.3376502
[60] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2020. Drawing from justice theories to support targets of online harassment. *New Media and Society* (2020), 1–23. https://doi.org/10.1177/1461444820913122
[61] Sarita Schoenebeck, Carol F Scott, Ellen Selkie, Emma Hurley, and Tammy Chang. 2021. Youth Trust in Social Media Companies and Expectations of Justice: Accountability and Repair after Online Harassment. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021). https://doi.org/10.1145/3449076
[62] Joseph Seering. 2020. Reconsidering Community Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Hum.-Comput* 107 (2020), 28. https://doi.org/10.1145/3415178
[63] Joseph Seering, Tianmi Fang, Luca Damasco, Mianhong Cherie Chen, Likang Sun, and Geoff Kaufman. 2019. Designing user interface elements to improve the quality and civility of discourse in online commenting behaviors. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, 14. https://doi.org/10.1145/3290605.3300836
[64] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2020. Metaphors in moderation. *New Media & Society* (2020), 146144482096496. https://doi.org/10.1177/1461444820964968
[65] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on Twitch through moderation and example-setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. 111–125. https://doi.org/10.1145/2998181.2998277
[66] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media and Society* (2019), 1–28. https://doi.org/10.1177/ToBeAssigned
[67] Kay Kyeongju Seo. 2007. Utilizing peer moderating in online discussions: Addressing the controversy between teacher moderation and nonmoderation. *American Journal of Distance Education* 21, 1 (5 2007), 21–36. https://doi.org/10.1080/08923640701298688
[68] Leif Singer, Fernando Figueira Filho, Brendan Cleary, Christoph Treude, Margaret Anne Storey, and Kurt Schneider. 2013. Mutual assessment in the social programmer ecosystem: An empirical investigation of developer profile aggregators. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. 103–116. https://doi.org/10.1145/2441776.2441791
[69] Tim Squirrell. 2019. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media and Society* (2019). https://doi.org/10.1177/1461444819834317
[70] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators. In *Proceedings of 2021 CHI Conference on Human Factors in Computing Systems - CHI '2021*. ACM, 14. https://doi.org/10.1145/3411764.3445092
[71] Daniel J Steinbock. 2005. Data Matching, Data Mining, and Due Process. *Georgia Law Review* 40, 1 (2005), 3–84.
[72] Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. 2010. A combination approach to web user profiling. *ACM Transactions on Knowledge Discovery from Data* 5, 1 (2010), 44. https://doi.org/10.1145/1870096.1870098
[73] Thomas Theodoridis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2015. Assessing the reliability of facebook user profiling. In *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web*. 129–130. https://doi.org/10.1145/2740908.2742728
[74] Catalina L. Toma, Jeffrey T Hancock, and Nicole B Ellison. 2008. Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin* 34, 8 (2008), 1023–1036. https://doi.org/10.1177/0146167208318067
[75] Twitch.tv. 2020. *Transparency Report 2020.* Technical Report. https://www.twitch.tv/p/en/legal/transparency-report/
[76] Emily van der Nagel and Jordan Frith. 2015. Anonymity, pseudonymity, and the agency of online identity: Examining the social practices of r/Gonewild. *First Monday* 20, 3 (2 2015). https://doi.org/10.5210/fm.v20i3.5615
[77] Ting Yu Wang, F Maxwell Harper, and Brent Hecht. 2014. Designing better location fields in user profiles. In *Proceedings of the ACM Conference on Supporting Group Work*. N Y, USA, 73–80. https://doi.org/10.1145/2660398.2660424
[78] Michael Wenzel, Tyler G. Okimoto, Norman T. Feather, and Michael J. Platow. 2008. Retributive and restorative justice. *Law and Human Behavior* 32, 5 (2008), 375–389. https://doi.org/10.1007/s10979-007-9116-6
[79] Donghee Yvette Wohn. 2019. Volunteer moderators in Twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of 2019 ACM Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3290605.3300390

[80] Donghee Yvette Wohn and Guo Freeman. 2020. Audience Management Practices of Live Streamers on Twitch. In *IMX 2020 - Proceedings of the 2020 ACM International Conference on Interactive Media Experiences*. Association for Computing Machinery, Inc, New York, NY, USA, 106–116. https://doi.org/10.1145/3391614.3393653

[81] Yu Chu Yeh. 2010. Analyzing online behaviors, roles, and learning communities via online discussions. *Educational Technology and Society* 13, 1 (2010), 140–151. http://www.jstor.org/stable/pdf/jeductechsoci.13.1.140.pdf

[82] Amy X. Zhang and Justin Cranshaw. 2018. Making sense of group chat through collaborative tagging and summarization. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 27. https://doi.org/10.1145/3274465

[83] Leizhong Zhang, Qiong Yang, Ta Bao, Dave Vronay, and Xiaoou Tang. 2006. Imlooking: Image-based face retrieval in online dating profile search. In *Companion of the ACM Conference on Human Factors in Computing Systems*. 1577–1582. https://doi.org/10.1145/1125451.1125739

[84] Chen Zhao and Gonglue Jiang. 2011. Cultural differences on visual self-presentation through social networking site profile images. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 1129–1132. https://doi.org/10.1145/1978942.1979110

## A APPENDIX: PARTICIPANT TABLE

Table 1. Demographic and Experience of Participants

| ID | Option | Viewership | Category | Experience (yrs) | Age | Race | Gender |
|---|---|---|---|---|---|---|---|
| P1 | A | 18-20 | Gaming | 4 | 21 | Hispanic | F |
| P2 | A | 10-15 | Gaming | 4 | 19 | African American | M |
| P3 | A | 70-100 | Art, body painting | 2.5 | 23 | Hispanic | M |
| P4 | B | — | Gaming | 1.5 | 18 | White | M |
| P5 | B | — | Gaming | 3 | 27 | African American | F |
| P6 | B | — | Gaming | 3.5 | 34 | White | F |
| P7 | A | 30-35 | Rhythm & music game | 0.5 | 18 | White | M |
| P8 | A | 15-20 | Gaming, video editing | 4 | 18 | White | F |
| P9 | B | — | Gaming | 1 | 19 | White | M |
| P10 | A | 130-150 | Gaming | 2 | 18 | Asian | M |
| P11 | B | — | Gaming | 3 | 19 | White | F |
| P12 | A | 650-1400 | Gaming, IRL | 2 | 21 | Asian | M |
| P13 | B | — | Gaming, IRL, Drama | 3 | 29 | White | M |
| P14 | B | — | Gaming, IRL | 8 | 28 | White | F |
| P15 | A | 800-1000 | Gaming, IRL, eSports | 6 | 31 | White | M |
| P16 | B | — | Gaming, IRL | 3 | 24 | Pacific Islander | M |
| P17 | A | 9000-11000 | Gaming | 1.5 | 21 | White | M |
| P18 | A | 3000-4000 | Gaming | 1 | 20 | White | F |
| P19 | B | — | Gaming | 5 | 26 | Asian | M |

January 2021 (submitted), April 2021 (revised), July 2021 (revised), July 2021 (accepted).