Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting

Joseph Seering
Human-Computer Interaction Institute
Carnegie Mellon University
jseering@cs.cmu.edu

Robert E. Kraut
Human-Computer Interaction Institute
Carnegie Mellon University
robert.kraut@cmu.edu

Laura Dabbish
Human-Computer Interaction Institute & Heinz College
Carnegie Mellon University
dabbish@cs.cmu.edu

ABSTRACT

Online communities have the potential to be supportive, cruel, or anywhere in between. The development of positive norms for interaction can help users build bonds, grow, and learn. Using millions of messages sent in Twitch chatrooms, we explore the effectiveness of methods for encouraging and discouraging specific behaviors, including taking advantage of imitation effects through setting positive examples and using moderation tools to discourage antisocial behaviors. Consistent with aspects of imitation theory and deterrence theory, users imitated examples of behavior that they saw, and more so for behaviors from high status users. Proactive moderation tools, such as chat modes which restricted the ability to post certain content, proved effective at discouraging spam behaviors, while reactive bans were able to discourage a wider variety of behaviors. This work considers the intersection of tools, authority, and types of behaviors, offering a new frame through which to consider the development of moderation strategies.

Author Keywords

Moderation strategies; chatroom behavior; authority and imitation.

ACM Classification Keywords

H.5.3 Group and Organization Interfaces: Synchronous Interaction

INTRODUCTION

In thriving online communities, a rough consensus generally emerges about norms, i.e. the range of acceptable behaviors. Norms of appropriate behavior vary substantially across communities. Personal insults may be the primary way to interact in one community, but may be frowned upon in another. Wikipedia expects writers to adopt a neutral point of view when writing articles, while the Huffington Post expects guest bloggers to express a viewpoint. PsychCentral.com, a site with more than 150 health support communities, prohibits conducting any type of research on the site for publication or educational purposes [42]. Snapchat users share mundane everyday moments in solidarity with friends [9].

Violations of these standards of appropriate behavior often undermine a community's purpose and drive members away. Sexual harassment of women in online games causes them to try to hide their identity or to leave the game entirely [19]. Unwanted sexual solicitation on social networks damages the potential of these spaces for socializing [54]. Facebook memorial page trolls disrupt the process of grieving after the death of loved ones [41].

Anonymous and pseudonymous online communities are particularly challenging to regulate. Without any easy way to attribute user behavior to real-life identities, users can behave in virtually any way they like without fear of reprisal or loss of reputation [16][44][50]. Within anonymous and pseudonymous online communities, a wide variety of behaviors are observed including vicious trolling and harassment as well as supportiveness and strong empathy [13][28][31]. This sort of extreme variation can happen even across the same platform in different channels or forums, which suggests that small differences between these spaces can have substantial impact on behavior. These technological variations intersect with community needs and goals to construct norms [10][14][34][51].

Communities can use formal and informal methods to enforce standards of appropriate behavior [30], including explicit rules, reputation systems that provide incentives for people to act appropriately, methods to report inappropriate behavior, and algorithms that automatically remove offending behavior. While these approaches help deal with misbehavior, anti-normative behavior is still a substantial problem on a variety of platforms [5].

This paper explores how moderation tools and imitation effects can address norm violations. We explore the effectiveness of both proactive (preventive) and reactive (punitive) moderation tools, where proactive tools prevent certain behaviors while reactive tools punish users after-the-fact for engaging in them. We also explore the potential

CSCW '17, February 25-March 01, 2017, Portland, OR, USA
© 2017 ACM. ISBN 978-1-4503-4335-0/17/03…$15.00
DOI: http://dx.doi.org/10.1145/2998181.2998277

for authority figures to shift culture by modeling positive behaviors other users can emulate.

This paper builds on previous research on imitation and deterrence and applies these concepts to moderation in a pseudonymous, ephemeral environment. First, building on lessons from theories of imitation [6][11][12][15][38][53], we show that text chat behaviors on Twitch (twitch.tv), a video streaming platform, are contagious, and that anti-social and pro-social behaviors spread differently. We show that the behavior of individuals with authority in and commitment to a particular channel has greater influence over the behavior of others, but that "outsider" users without specific status in a channel have no additional impact. Second, drawing from Deterrence Theory [17][25][37][49], we demonstrate that different approaches to moderation can reduce the spread of different types of behavior. Setting a chatroom to a restrictive mode reduces the frequency of spam overall while having no substantial effect on other types of behavior. Banning a user after they post a message of a certain type lowers rates of that type of message in subsequent posts.

BACKGROUND

This paper explores two approaches to influencing behaviors on Twitch: behavioral imitation, where observing one type of behavior encourages observers to behave in the same way, and deterrence, where threat of punishment or enacted punishment causes users to change their behavior. This section explores relevant findings in the literature in each of these mechanisms, and considers what each theory would predict for the Twitch context.

Imitation and Conformity

Several related processes lead to the spread of behavior from one person to another. Theories of imitation discuss how individuals learn to behave and think in ways that are similar to their peers. Theories of obedience and conformity describe the power of authority figures to influence behavior.

Imitation occurs in two primary phases. The initial effect of observing others' behaviors, as described by Wheeler, is in helping the individuals conceive of those behaviors as possible courses of action when they might not otherwise have considered them [53]. Here we will refer to this as the conception effect. Second, when choosing from a set of possible behaviors, individuals are more likely to choose the behavior that their peers prefer or which they perceive to be in accordance with social norms [6][12][15][38]. The more peers observed acting in a certain way, the more likely the individual is to do so as well, independent of whether such an action is a good idea.

Studies have found the presence of imitation effects in a variety of different contexts, from the mimicry of non-verbal behavior and language [11], to expressions of attitudes and beliefs [7][48] to speeding [46], to copycat suicide [40].

Social learning theory contributes an additional related perspective on mechanisms involved in the adoption of deviant and conforming behavior [3][24]. This framework explains a variety of what Akers describes as deviant behaviors including adolescent marijuana usage [1], teenage cigarette smoking [4], and alcohol use among the elderly [2].

We hypothesize that behavior on Twitch can be explained by these same types of imitation effects. Rare behaviors will be imitated at particularly high rates as a result of the conception effect [53]; while on Twitch certain behaviors such as spam are very common and users need no reminder that they are possible courses of action, reminders to be polite and kind will have more influence due to the relative rarity of those types of behaviors.

H1: An instance of a given behavior on Twitch will increase the likelihood of other users engaging in that behavior.

Literature on conformity and obedience to authority supplements the perspective provided by imitation theories. Milgram's classic shock experiments showed how powerful the effect of an authority figure can be on overriding individual tendencies [35]. Nurses were willing to follow alarmingly bad directions if encouraged to do so by an authoritative doctor [26]. Psychological journal articles were much more likely to be accepted for publication if they had well-known researchers' names attached [39]. Overall, people are more likely to follow the example or take the word of others who have explicit authority.

Imitation has also been documented in Human-Computer Interaction literature [8][29][45][55]. Bakshy et al. [8] found that Facebook users were significantly more likely to share a link if their friends had shared it. This effect was limited to a brief period of time after they saw their friends' posts. Zhu, Kraut, and Kittur also found significant imitation and conformity effects among Wikipedia workers collaborating on projects, though this effect varied significantly based on level of users' identification with the group [55].

In each of the cases discussed above, participants were more likely to do something that they would otherwise be uncomfortable doing, or that under other circumstances they would find reprehensible. From conformity theories we draw the hypothesis that users with authority or status will be more likely to be emulated:

H2: Users with more authority or status within a community will be imitated with greater frequency.

There are four different status badges that users in this dataset could have in a given channel, each of which was denoted by a specific badge that appeared next to their name when they posted in chat. *Channel owners* were users who broadcast the content on the channel. They had the highest status, and had ultimate authority over the various moderation tools. *Moderators* were users designated by the channel owner to enforce behavioral standards through bans and chat moderation modes. They had authority and status within the community. *Subscribers* were users who paid a monthly fee of approximately $5, of which part went to the channel owner and part to Twitch, to gain special privileges in a channel and to support the broadcaster. Subscription is only available as an option in sufficiently large or well-known channels, but through subscription to a channel a user may quickly progress to a higher status by way of demonstrated commitment to the channel [21]. *Twitch Turbo* users were users who paid approximately $9 per month for a premium account to gain small benefits across the whole site, including removal of ads. In this dataset, a Twitch Turbo badge was a mark of some status, but not status within any specific ingroup. In this study, we will refer to users with none of these statuses as regular users. Note that we do not include channel owners' chat messages in our analyses because channel owners usually communicate with viewers by speaking directly as part of their broadcast.

Here we use these user status categories as representative of different levels of authority and commitment to a channel. Per Hogg's social attraction hypothesis [27], ingroup members are more liked and thus more influential because they are perceived as conforming to a positive ingroup prototype. Here, users who like or want to be part of a particular channel community see the behavior of subscribers as clear examples of ingroup norms, as these subscribers have demonstrated explicit loyalty to the channel. Moderators, through both their additional abilities and their status as a favorite of the channel owner, exemplify Hogg's prototypical leaders: they have disproportionate power to determine standards of conduct, define identity, and organize and guide discussion. In contrast, Twitch Turbo users had a literal status icon next to their names, but this status was not specific to the channel. Within the channel they are as much outsiders as regular users if not more so.

Deterrence Theory

Where imitation and social learning theories focus primarily on the ways behaviors are learned from peer groups and networks, Deterrence Theory focuses specifically on the impact of punishment on deterring certain types of behaviors, and is used here as a reference for understanding the impact of certain moderation strategies on behaviors [17]. Deterrence Theory distinguishes between general and specific deterrence, where specific deterrence is defined as the impact of punitive actions on individuals upon which they are enforced, and general deterrence is the impact of the threat of such action on uninvolved observers. Deterrence theory has also been conceptualized as indirect vs direct experiences with punishment [49]

The theory of general deterrence suggests that the threat of arrest and punishment may deter criminals from committing crimes, and that different levels of certainty and severity of punishment will affect the effectiveness of this deterrence. This effect has been demonstrated in online contexts [25]. Nagin [37] identifies several methods for studying deterrence in the wild, including interrupted time-series studies, ecological studies, and perceptual studies. This current study uses the interrupted time-series method, a quasi-experimental method that looks at differences in behaviors immediately before and after an intervention. In this case we look at chat behaviors prior to and subsequent to a deterrence event that we hypothesize will affect these behaviors and compare the effects of deterrence on different categories of behaviors. By looking at thousands of instances of the intervention, our procedure guards against many confounds associated with interrupted time series methods, e.g. other historical events occurring simultaneously with the intervention.

The Twitch platform offers several tools to help moderate offensive behaviors. Broadcasters and moderators can ban users directly for variable lengths of time in response to an offensive or inappropriate message. Various third party chat-moderation bots can also be installed to automatically ban users who post certain specified types of content. Channel chat moderation modes can be enabled by channel owners or moderators proactively to prevent certain types of posting behavior.

While Nagin [37] discusses the challenges of understanding the connections between perceptions of abstract policies and impact on an individual's behavior, the immediacy of punishment on Twitch helps avoid this particular pitfall. Where real-world legislators may seem distant from the corporeal behavior that they regulate, the "policy-makers" on Twitch are often literally visible to their audience. Furthermore, punishments in the form of bans are relatively common and are visible to all users, so participants have clear and direct evidence of what is and is not considered appropriate behavior. In many cases they are often even told directly by chat moderation bots or human moderators what specific behavior caused the ban.

We hypothesize the presence of a generalized preventative effect from proactive moderation techniques on Twitch, which take the form of chat moderation modes. Channel chat moderation modes are tools available on all channels that restrict users' posting behaviors. The three modes we explored are subscribers-only mode, where only subscribers to the channel may chat, slow-mode, where users have to wait a specified amount of time between sending messages, and R9k-beta mode, where users are prohibited from posting lengthy content that has already been posted. Whereas traditional bans are imposed in response to messages, these modes prevent messages from being posted at all except under the designated circumstances. These modes cannot be customized to target different types of unwanted behavior, beyond the choice to enable or disable them in response to different chatroom conditions. As such, channel owners and moderators cannot directly customize their usage to encourage particular behaviors; these modes only serve to make it more difficult to engage in specific anti-social behaviors, namely spam. Because of this, they will have a generalized preventative effect on anti-social behavior, but will not exhibit classic generalized deterrent effects and will not directly encourage pro-social behavior.

H3: When chat moderation modes are enabled, the frequency of spam will decrease, but the frequency of other more prosocial behaviors will not increase.

In contrast to the limited customizability of chat moderation modes, bans are completely customizable. The owner of a particular channel chatroom and the moderators they designate choose which users to ban and for how long, and what settings to use on moderation bots that ban users by proxy. Bans, like chat moderation modes, will have a deterrent effect, but this effect will be more flexible and will deter different types of behavior depending on how bans are applied.

H4: When a particular type of behavior is banned, the subsequent messages will have a lower frequency of this type of behavior.

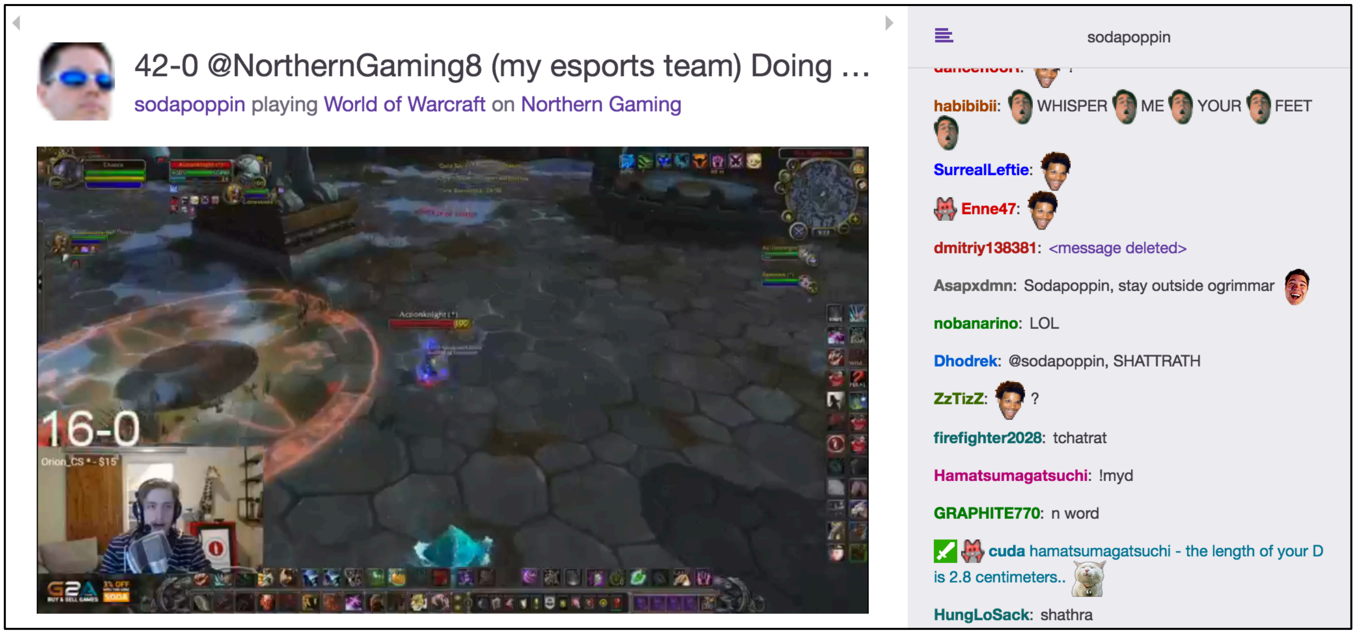
While we have framed this hypothesis in terms of deterrence theory, imitation theory makes a similar prediction, although for a different reason. Banning a particular type of content removes it from the view of other users, who then simply may not conceive of it as possible behavior [53].

TWITCH PLATFORM DESCRIPTION

Twitch is a video-streaming website where users can broadcast live video of themselves engaged in various activities to viewers, with whom they can interact via webcam or typing in a chatroom attached to the channel. Hamilton, Garretson, and Kerne [21] describe Twitch channels as participatory communities, where users engage in banter and conversation according to an established but continuously evolving set of norms. See Figure 1 for an example of a Twitch channel. With more than 16 million unique visits monthly [43], Twitch commanded the fourth-highest proportion of peak internet traffic by data volume in 2014, behind only Netflix, Google, and Apple [33]. Many thousands of video streams are active at any given time on Twitch, with most based around video and computer gaming and a small fraction based around creative endeavors such as painting or playing music.

Both broadcasters and viewers on Twitch use pseudonyms, and a wide variety of behaviors and behavioral norms can be found on different channels. As in many anonymous and pseudonymous communities, trolls can be a disruptive presence [23]. While behavioral norms vary across different channels, behaviors that are generally considered to be disruptive include spamming capital letters or emotes, harassing other viewers or the broadcaster, posting links to explicit content or malware, or starting conversations about incendiary topics [18].

Twitch is an ideal platform on which to study the effects of different moderation techniques. With thousands of channels with different numbers of viewers, different approaches to moderation, and different behavioral norms, analysis of chatroom data allows for a better understanding of what works to stop the spread of undesired behaviors. Behavioral imitation is very visible on Twitch. For example, users in larger channels often engage in spamming of “copypasta,” which are long, often-nonsensical messages with many emotes that users copy and paste into a chat repeatedly. While such behavior may be desirable on some channels and in some cases it can even be compared to the type of cheering that happens at sporting events [22], many broadcasters prefer to keep their chat rooms civil and thus seek tools that will stop the spread of such behavior when it appears.



42-0 @NorthernGaming8 (my esports team) Doing …

sodapoppin playing World of Warcraft on Northern Gaming

sodapoppin

habibibii: WHISPER ME YOUR FEET

SurrealLeftie:

Enne47:

dmitriy138381: <message deleted>

Asapxdmn: Sodapoppin, stay outside ogrimmar

nobanarino: LOL

Dhodrek: @sodapoppin, SHATTRATH

ZzTizZ: ?

firefighter2028: tchatrat

Hamatsumagatsuchi: !myd

GRAPHITE770: n word

cuda hamatsumagatsuchi - the length of your D is 2.8 centimeters..

HungLoSack: shathra

Figure 1: Sample Twitch Channel Interface

PROCEDURE

This study involved the collection and analysis of approximately 21 million messages sent to channel chatrooms on Twitch over the course of one week in early March 2016. In this section, we first describe the method for gathering data and then describe the four analyses performed on this data. The first analysis demonstrates clear imitation effects for three different types of behavior that we studied: spam, questions, and smiles. The second analysis shows that the status and authority of users within the ingroup affects the amount that they are imitated, and that the effects vary across different behaviors. The third analysis shows differential results from reactive and proactive approaches to moderation. Finally, the fourth analysis takes a different approach and attempts to determine the duration of the impact of these effects by looking at the strength of the imitation effect over time.

Data collection

In this study we identified three categories of behavior that represented different modes of interaction on Twitch in order to understand how they spread and persisted or disappeared in response to moderation. First, we defined "spam" using the default settings on one of the most widely used Twitch chat moderation bots as an example of anti-normative behavior. By this definition, spam messages were those with a large number of emotes, capital letters, or symbols. Second, we identified conversational messages, where users ask questions of the broadcaster or each other as an example of neutral behavior. These are messages that end with a question mark. Third, we identified messages with positive emotions, which in this case are defined as those with positive smile emoticons, as an example of pro-social behavior. Note that we classify large numbers of emotes unique to Twitch as spam in accordance with common moderation bot rules and with the understanding that many of these emotes are used for trolling [19], but we found in our analysis that single uses of traditional smiley-face emoticons were almost always positive or only lightly teasing.

Note that for this study we focused on Twitch-wide emotes instead of channel-specific emotes; the Twitch-wide emotes are much more widely-used, and meaning can be generalized across channels. "Kappa", the most used emote on Twitch, is posted upwards of one million times per day across Twitch [20].

We selected these categories as examples of a range of behaviors, but it is important to note that in many channels on Twitch "spammy" messages are tolerated or even encouraged. Table 1 shows the three types of messages that were analyzed and their overall frequency in the dataset.

Type	Criteria for inclusion	Action Valence	Overall Frequency as % of Messages
Spam	Many emotes, capital letters, or symbols	Anti-normative	14.8%
Questions	Ends with "?"	Neutral	4.7%
Smiles	Contains ":)", ":D", ":P", or ";)"	Pro-social	0.9%

Table 1: Categories of Messages

Contagion of the sort that we discuss here can take a variety of forms. For example, spam contagion might be started by a small number of users to disrupt a particular conversation:

> **USER1:** go [play] constructed please!
> **USER2:** constructed is usually boring to me
> **USER3:** I wouldn't mind watching you do maybe a couple of hours of constructed a week, just for a change.
> **USER4:** are you ever going to play SMITE again @[streamer]?
> **USER5:** VapeNation
> **USER6:** V/\ Vape Nayshon!
> **USER7:** NapeVation /\V
> **USER8:** why are all u idiots spamming vapenation
> **USER9:** I love waiting for a comment to get [banned] haha
> **USER10:** VapeNation
> **USER11:** VapeNation V/\
> **USER12:** VapeNation V/\

In this example, users were having a conversation with the streamer about what the streamer should do next when they were interrupted by other users spamming variants of "VapeNation", a popular meme-phrase referring to smoking with a vaporizer. Regular users expressed irritation about this disruption and requested bans for the spammers, but in this case the moderators did not intervene and the spamming continued for another five minutes before it tapered off.

Imitation of question-asking behaviors might be encouraged by a perceived likelihood to get a response from the streamer or a moderator:

> **USER1:** @[streamer] what do u think about jax jungle ?
> **USER2:** @[streamer] Any chance we can see Skarner jng of the team comps permit it?
> **USER3:** @[streamer] What do you think of Aurelion Sol? What's the best role to play him/her in?
> **USER4:** @[streamer] just got out of a game with morgana..bot.we were ok...adc was feded..but mid and jungle fed...so what do i do in such occasions??

The streamer in this example, who is a highly ranked player of the popular game League of Legends, probably responded verbally in the stream to the first question asked here, so other users decided that it was a good time for them to ask questions as well. Note that this mechanism is imitation but is mediated by the actions of a third party, which in this case is the streamer.

Finally, imitation of smiles often comes from positive exchanges between users in the chat, and can spread to other interactions as well:

> **USER1:** hiii @[USER2] :D
> **USER2:** @[USER1], hey :)
> **USER1:** :D
> **USER2:** hi chat :)

Here one user greets another in a friendly way, and this user responds positively and in turn greets the other users.

While most Twitch users only see the graphical interface for the channel chatrooms (see Figure 1), the underlying structure of the chats is based on IRC (Internet-Relay-Chat) protocols. We created an IRC "chatbot" in Python that uses the Socket module for the purpose of collecting messages and relevant metadata from channels across Twitch. The data collection process involved three steps: determining which channels to scrape, creating the data collection scripts, and running the scripts over the period of one week to create the dataset. This chat scraping was done with the permission of data science staff at Twitch.

This study used sample a of roughly 600 Twitch English-language channels stratified by size. In order to select channels for analysis, we created ten strata of channels, based on their number of viewers, where each stratum contained approximately the same number of viewers. The first group, which contained the largest channels, had the fewest channels overall, while each successive group had more channels with fewer viewers. We randomly selected 100 channels from each of these groups with the exception of the first group and the third group, which contained only 81 and 46 channels total.

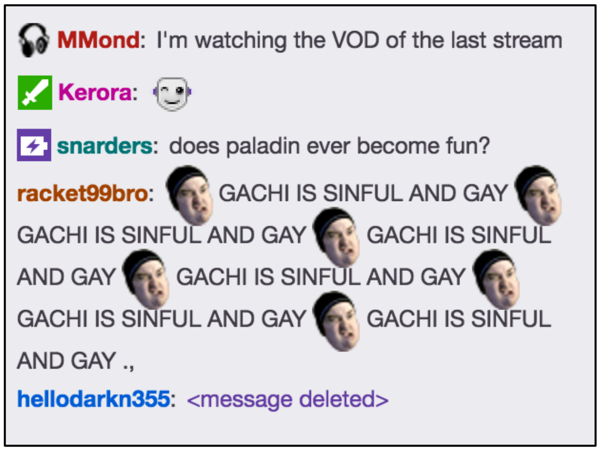
We ran the message collection script for nine days in early March 2016. During this time period, every message sent to one of the channels in our sample was recorded. Overall, approximately 21 million messages were collected from English-speaking channels.

Features of the dataset

Each message in the dataset was tagged with several features. These included the channel name, username, timestamp, and message as well as whether the user who sent the message was a moderator, subscriber, or Twitch Turbo user. We removed usernames from our dataset prior to analysis. We also tracked whether each channel had a channel chat moderation mode set while any given message was sent, which would have influenced the types of messages that could be sent. While messages were not directly tagged as being banned, as bans are user-specific not message-specific, we inferred that a specific message was banned if it was the most recent message sent by a user who was banned, under the assumption that users are most frequently banned in response to the most recent message they sent prior to their ban.

After the data was collected, we tagged each message with a variety of characteristics based on the text of the message. These tags included flags for each of the three types of behavior described above: spam, questions, and smiles. Each message was also tagged with the number of messages being sent in the channel per second at the time it was sent, which was used as a measure of how many users were actively chatting in the channel at any given time. Table 2 shows the variables used in this study.

Figure 2 shows examples of each type of user and each type of content. The first message in Figure 2 has none of the content types we flagged, and was posted by a subscriber. Each subscriber icon is unique to the channel to which it is attached; in this case, the subscriber icon for a music channel is a set of headphones. The second message is a smile posted by a moderator, indicated by a green sword icon. The third message is a question posted by a Twitch Turbo user, as indicated by a purple battery icon. The last two messages were posted by regular users. The first of these has spam with many capital letters and emotes, and the second is marked as deleted because of a ban.



MMond: I'm watching the VOD of the last stream

Kerora:

snarders: does paladin ever become fun?

racket99bro: GACHI IS SINFUL AND GAY GACHI IS SINFUL AND GAY GACHI IS SINFUL AND GAY GACHI IS SINFUL AND GAY GACHI IS SINFUL AND GAY GACHI IS SINFUL AND GAY .,

hellodarkn355: <message deleted>

Figure 2: User Types and Behaviors

In order to analyze the impact of specific events on the messages that followed them, messages were grouped into clusters of twenty-one for all analyses except the analysis of moderation mode. These clusters contained ten messages to establish the state of the channel prior to the an event of interest, one "event" message with properties that served as independent variables, and ten messages following the "event" with characteristics that served as dependent variables. We included all messages as events that had enough messages sent before them and after them to use in analysis. In the corpus of approximately 21 million messages, we analyzed approximately one million event messages that had ten messages prior and ten subsequent to each. In the moderation mode analysis we analyzed groups of forty messages: twenty before an event and twenty after an event. In this case the event in question was a change in chat moderation mode that occurred in between the twentieth and twenty-first messages.

The "Rate" variable used in this analysis, which notes how many messages per second were being sent in a given chatroom at the time of the event, is used here as a proxy for size of channel. We used rate as measure of size instead of number of viewers because different channels may have higher or lower rates of participation, and total viewer numbers can be skewed by tools like viewbots, which are used to artificially inflate the number of viewers in a channel. In contrast, the rate of messages being sent at a given time models the user's perception of the size and level of activity of the crowd. Across all analyses, a higher rate of messages being sent was associated with more spam, fewer questions, and fewer smiles.

These analyses all use interrupted time-series models as described in Nagin [37]. Shadish, Cook, & Campbell [47] describe two major challenges in proving causality. First, when it is argued that an event A caused an event B, it must be shown that the reverse explanation could not be true (i.e. B did not cause A). The interrupted time-series approach resolves this problem by analyzing events that occur in fixed temporal sequences; B could not have caused A because B occurred after A.

Variable Name	Description	Mean
PriorState	Number of messages of a given type (spam, questions, smiles) in the past ten messages. Sum of Boolean for each of the ten prior messages. Centered at mean = 0.	Spam: 1.48 Questions: 0.47 Smiles: 0.09
Event	Whether a given message contains spam/question/smile. 1 = YES, 0 = NO	Spam: 0.148 Questions: 0.047 Smiles: 0.009
Rate	Messages per second being sent in the chat at the time of the event. Centered at mean = 0.	4.9
isMod	Whether a given message was sent by a moderator. 1 = YES, 0 = NO	0.083
isSub	Whether a given message was sent by a subscriber. 1 = YES, 0 = NO	0.235
isTurbo	Whether a given message was sent by a Turbo-user. 1 = YES, 0 = NO	0.036
isBanned	Whether a given message was banned. 1 = YES, 0 = NO	0.021
R9k	Whether a given message was sent in R9kbeta mode. 1 = YES, 0 = NO	0.154
Slow	The number of seconds of slow mode currently on. If slow mode is off, equals zero.	13.82
Sub	Whether a given message was sent in subscribers-only mode. 1 = YES, 0 = NO	0.028

Table 2: Variables, Descriptions and Means

The second challenge to proving causality is disproving the possibility of alternative explanations based on outside events. In this case, the most plausible outside explanation is that behavior displayed on the video stream, which was not captured in this dataset, explains the relationship between event messages and subsequent chat behavior. This possibility will be explored in more depth in the final section of this paper, but it does not conflict with the implications of this work.

All data in this study was explored in aggregate, and no individual message text was analyzed by the researchers beyond scripted tagging of messages according to the three content types. Messages in the dataset were tied to users only through pseudonymous usernames, which were eliminated from the final dataset used for analysis. No experiments were performed, and no users encountered a different experience on the website than they normally would have.

ANALYSIS 1: EFFECTS OF IMITATION

Before attempting to understand the impact of key individuals and moderation techniques on imitation, we need to first demonstrate that imitation occurs in this setting. Thus to test H1, we analyzed the impact of "event" messages on subsequent content.

Table 3 shows the regression coefficients for a linear regression on percentage of a given type of behavior in the ten messages following an event as a function of properties of this event, prior state, and rate of messages at the time of the event. Note that the PriorState and Rate variables were centered to have a mean of zero, which causes the intercept to be equal to the percentage of messages of the given type in the next ten messages in a channel with average characteristics when the "event" message has none of the possible characteristics. This linear regression followed the format:

P' = Intercept + PriorState + Event + Rate

Using the coefficients listed in Table 3, the percentage of messages containing spam in the next ten messages following an event is 13.7% + 5.7% times the number of spam messages above the average amount of spam in the prior ten messages plus 6.0% if the event message was spam, plus 0.5% per additional message per second being sent at the time of the event above the average rate.

In each case, regardless of whether an event message contained spam, questions, or smiles, the following ten messages had a significantly higher proportion of messages of that same type than if the event message had not been of that type, after controlling for rate of messages in the chat and prior quantity of each type of behavior. Of these, smiles were imitated significantly more than either other type. This confirms H1, as these behaviors display imitative properties. Our results suggest that rare behaviors may be imitated with a greater frequency, but we did not have enough categories of behavior in our analysis to state this conclusively.

	Dependent Variables:					
	Percentage Spam		Percentage Questions		Percentage Smiles	
Independent Variables:	**Coefficient**	**SE**	**Coefficient**	**SE**	**Coefficient**	**SE**
Intercept	13.7%***	0.01%	4.7%***	0.00%	1.0%***	0.00%
PriorState	5.7%***	0.01%	2.3%***	0.01%	2.0%***	0.01%
Event[1]	6.0%***	0.03%	2.6%***	0.03%	2.2%***	0.03%
Rate	0.5%***	0.00%	-0.3%***	0.00%	-0.7%***	0.00%
N=2032436	$R^2 = 0.44$		$R^2 = 0.11$		$R^2 = 0.06$	

* p < 0.05 ** p < 0.01 *** p < 0.001

Table 3: Impact of a given type of "event" message on the subsequent ten messages

[1] "Event" is a binary variable noting whether the event message contained the type of content being analyzed in the given column. For example, event would equal one in the left column if the message contained spam, one in the middle column if the message contained a question, and one in the right column if the message contained a smile. In this case, for example, if the event message contained spam, an additional 6% of the subsequent ten messages contained spam.

Table 4 shows the percentage increase in each type of behavior following the event. For example, when an event message contained spam, we observed 43.8% more messages containing spam over the next ten messages. Of the three types of messages, smiles were most susceptible to these imitation effects and spam was the least susceptible.

Behavior	% Increase after Event
Spam	43.8%
Question	55.3%
Smile	220.0%

Table 4: Percentage Increase after Event of Same Type

Discussion

The above analyses provide strong evidence for the presence of imitative effects in Twitch chatrooms, supporting H1. New users may be introduced to various types of behaviors by observing others engaging in them. Existing users may be reminded about particular behaviors or encouraged to engage in them when other users do so. These effects can be seen as a combination of the conception effect [53], where users are reminded of the possibility of engaging in a particular behavior, and broad imitation effects. In this case the most common behavior showed the smallest increase in percentage as a result of an imitation effect. This may be the result of small impact of a conception effect; because spam is so common, users do not need to be reminded that they can post spam themselves. Since smiles are quite rare, observing a smile reminds users of a mode of interaction that they might not otherwise have considered.

Of the above behaviors, questions are perhaps the least intuitively likely to be imitated, as each question expects a response that is not in the form of a question. One plausible scenario for imitation within the category of questions is that when a user receives a response to a question from the streamer or a moderator, other users ask questions because they believe they are likely to receive answers.

ANALYSIS 2: IMPACT OF USER-TYPE

In our second analysis, we further explored the imitation observed in Analysis 1 to test H2, that individuals with greater status and authority would be imitated more frequently.

We used four categories of users as examples of different levels of status and authority within a specific ingroup. Moderators show both status and authority; subscribers show status; regular users show neither commitment nor authority; and Twitch Turbo users show status, but not within a particular ingroup. In this sense, Twitch Turbo users can be compared to Wikipedia editors who are not attached to the specific project at hand, but still have some status [55]; Turbo users are not attached to the specific channel in which they are chatting, but they have a mark of status.

	Dependent Variables					
	Percentage Spam		Percentage Questions		Percentage Smiles	
Independent Variables	**Coefficient**	**SE**	**Coefficient**	**SE**	**Coefficient**	**SE**
Intercept	14.0%***	0.01%	4.8%***	0.01%	0.9%***	0.00%
PriorState	5.6%***	0.06%	2.3%***	0.01%	2.0%***	0.01%
Event[2]	5.4%***	0.03%	2.5%***	0.03%	2.1%***	0.03%
isMod	-0.7%***	0.04%	-0.04%*	0.02%	0.2%***	0.01%
isSub	-0.9%***	0.02%	-0.4%***	0.01%	0.04%***	0.01%
isTurbo	-0.4%***	0.05%	1.9%***	0.03%	0.03%**	0.01%
Rate	0.5%***	0.00%	-0.3%***	0.00%	-0.1%***	0.00%
Event*isMod	4.8%***	0.13%	1.2%***	0.12%	0.7%***	0.08%
Event*isSub	2.0%***	0.07%	0.07%	0.07%	-0.01%	0.06%
Event*isTurbo	-2.9%***	0.18%	-1.4%***	0.17%	-0.63%***	0.12%
N=2032436	$R^2 = 0.44$		$R^2 = 0.11$		$R^2 = 0.06$	

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 5: Impact of a message posted by a given type of user on subsequent messages of the same type

[2] "Event" is a binary variable noting whether the event message contained the type of content being analyzed in the given column. For example, event would equal one in the left column if the message contained spam, one in the middle column if the message contained a question, and one in the right column if the message contained a smile. In this case, for example, if the event message contained spam, an additional 5.4% of the subsequent ten messages contained spam.

In this analysis, as in the first analysis, groups of twenty-one messages were analyzed: ten messages establishing the prior state of the channel at the time of the event, one event message with characteristics that were treated as independent variables, and ten subsequent messages with characteristics that served as dependent variables.

Table 5 shows the regression coefficients for a linear regression on percentage of the given type of behavior in the ten messages following an event as a function of this event, prior state, rate of messages at the time of the event, type of user, and interaction between type of user and event. This linear regression followed the format:

*P' = Intercept + PriorState + Event + isMod + isSub + isTurbo + Rate + Event*isMod + Event*isSub + Event*isTurbo*

The results show that certain types of users had more impact than others. Status effects were persistent across all three categories of behavior, except in the case of Turbo users who had status but not within the ingroup.

Table 6 shows the impact of a given type of user posting a given type of content on percentage of that content over the subsequent ten messages. For example, when a moderator posted a message containing spam, 67.8% more messages with spam were posted in the next ten messages, while if a user with no authority posted a message with spam the number of messages with spam in the next ten messages increased by 38.6%.

Behavior	User Type	% Increase after Event
Spam	Mod	67.8%
	Sub	46.4%
	Turbo	15.0%
	Regular User	38.6%
Question	Mod	76.3%
	Sub	45.2%
	Turbo	62.5%
	Regular User	52.1%
Smile	Mod	333.3%
	Sub	236.7%
	Turbo	166.7%
	Regular User	233.3%

Table 6: Percentage Increase after Event of Same Type Posted by Given User Type

Discussion

Overall, different types of users were imitated at substantially different rates across the three categories of behaviors, supporting H2. This suggests that imitation and conformity to authority effects exist in this context.

Moderators were imitated significantly more than non-moderators when posting spam, questions, and smiles. Subscribers followed the same pattern as moderators, though they were significantly less influential in all three cases, and not statistically significantly influential on smiles or questions.

However, Twitch Turbo users, who paid monthly to have privileges across the site, were imitated less than regular users across all three categories of behavior. This could be explained by a perception of Turbo users as outsiders; because they openly display commitment to the whole site as opposed to one channel, they may be viewed as lacking commitment to the channel in which they are chatting. Zhu, Kraut, and Kittur [55] found that, among Wikipedia workers collaborating on projects, editors who identified with the projects acted similarly to prototypical group members, but editors who didn't identify with the group did not act similarly to prototypical group members at all despite their apparent status. This suggests that unless high status individuals also show commitment to the local community, members of the community will not imitate them.

Alternatively, Twitch Turbo users may feel that their Turbo subscription gives them some authority that other users do not think that they have. The data supports this second explanation; overall, users with only the Twitch Turbo tag were banned at approximately five times the rate of subscribers, who were almost never banned. While Twitch Turbo users may have some additional status as a result of their badge, regular users can plainly see that this badge does not afford them special treatment. More broadly, the results here provoke questions about how users with "premium" accounts like Twitch Turbo are viewed and what influence they buy with their subscription fees.

ANALYSIS 3: EFFECTS OF DETERRENCE

In the third portion of our analysis we explored the impact of moderation behaviors on subsequent message characteristics. Literature on deterrence suggests a number of ways that direct and indirect experiences with punishment might affect future behavior [32][52].

Impact of chat moderation modes on behavior

The first part of this analysis looks at behavior before and after a particular chat moderation mode was enabled in a chat in order to test H3, that chat moderation modes will deter spam but will not affect other types of behaviors.

By typing a command, channel owners and moderators can enable chat modes in a channel that restrict the types of messages that can be sent. Three chat modes were explored for this analysis: subscribers only mode, where only users who have subscribed to the channel may chat; slow mode, where users may only post messages every N seconds, where N is set by the moderator who enabled the mode; and R9K beta mode, where users who post messages of 9 characters or more are only permitted to post messages that have not previously been posted. This last mode is named after an experimental bot in a webcomic forum where it was first explored [36].

Category	Mode type	Change in next 20	Percent Change	P(Δ= 0)
Spam	Slow	-0.52	-14.3%	0.048*
	Sub	-0.95	-22.7%	0.003**
	r9k	-0.46	-14.7%	0.033*
Questions	Slow	0.16	14.6%	0.277
	Sub	-0.17	-18.9%	0.206
	r9k	0.27	31.3%	0.243
Smiles	Slow	0.08	38.4%	0.383
	Sub	0.10	66.4%	0.306
	r9k	0.08	66.1%	0.414

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 7: Percentage Change in Behaviors Following Implementation of a Chat Mode

These analyses were performed on groups of forty messages, where a change in channel mode occurred between the twentieth and twenty-first message in the set. The first twenty messages were compared to the second twenty messages to get an idea of channel behavior before and after the switch. The use of forty messages instead of twenty here is a response to the significantly smaller sample size; chat mode changes were relatively rare in this dataset. Slow mode, subscribers-only mode, and R9K beta mode were enabled 168 times, 667 times, and 97 times respectively in scenarios where there were twenty messages before and after the change during a single stream.

Overall, enabling each of these modes led to substantial decreases in spam in subsequent messages, but had no significant effect on subsequent behaviors of the other types. This supports H3.

Table 7 shows the impact of enabling chat modes on subsequent behavior. For example, after enabling subscribers-only mode there were 0.95 fewer messages containing spam in the next twenty messages, a decrease of 22.7%.

Impact of bans on subsequent imitation

To test H4, that bans would succeed in discouraging multiple types of behaviors we looked at the impact of banning a particular type of behavior on the frequency of that type of behavior in the next ten messages. We analyzed approximately two million groups of twenty-one messages. As in previous analyses, these groups consisted of ten messages prior to the event to establish a baseline for behavior at the time of the event, an event message with characteristics from which independent variables were drawn, and ten subsequent messages with characteristics that served as dependent variables. In approximately 2.3% of these cases, the event message was banned.

	Dependent Variables					
	Percentage Spam		Percentage Questions		Percentage Smiles	
Independent Variables	**Coefficient**	**SE**	**Coefficient**	**SE**	**Coefficient**	**SE**
Intercept	13.7%***	0.01%	4.7%***	0.01%	1.0%***	0.00%
PriorState	5.6%***	0.06%	2.3%***	0.07%	2.0%***	0.01%
Event[3]	6.4%***	0.03%	2.6%***	0.02%	2.2%***	0.03%
isBanned	0.1%	0.09%	0.6%***	0.04%	0.01%***	0.02%
Rate	0.5%***	0.00%	-0.3%***	0.00%	-0.1%	0.00%
Event*isBanned	-4.7%***	0.13%	-1.6%***	0.02%	-2.3%**	0.20%
N=2032436	$R^2 = 0.44$		$R^2 = 0.11$		$R^2 = 0.06$	

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 8: Impact of banning a particular type of message on subsequent messages

[3] "Event" is a binary variable noting whether the event message contained the type of content being analyzed in the given column. For example, event would equal one in the left column if the message contained spam, one in the middle column if the message contained a question, and one in the right column if the message contained a smile. In this case, if the event message contained spam, an additional 6.4% of the subsequent ten messages contained spam.

In nearly all cases, these data represent the impact of generalized deterrence – the indirect experience with punishment. Because users who were banned are typically prevented from posting for a period of time, other users typically posted the subsequent messages. These results represent the impact of observing another user being banned for a particular type of behavior.

Table 8 shows the regression coefficients for a linear regression on percentage of a given type of behavior in the ten messages following an event message as a function of event type, prior state, rate of messages at the time of the event, whether a message was banned, and interaction between ban and type of message. This linear regression followed the format:

*P' = Intercept + PriorState + Event + isBanned + Rate + Event*isBanned*

Table 9 shows the impact of bans on frequency of a particular behavior. Overall, banning any type of behavior had a significant negative impact on the frequency that behavior appeared in subsequent messages, confirming H4. For example, when a message containing a smile was banned, the percentage of smiles in the next ten messages decreased by 10.0%, but increased by 220.0% when the message was not banned.

Because bans did not necessarily occur directly after a message was posted, it is reasonable to question whether the amount of time between message posting and ban had an effect on subsequent messages. However, in our analyses we found no statistically significant impact. Given that the majority of bans in this dataset occurred within one second of message posting, chat moderation bots probably administered most of the bans automatically. In particular, in larger channels with higher rates of posting there was a higher proportion of rapid ban responses indicating more bot involvement. This suggests that broadcasters and moderators make use of available tools to administer bans before many more posts have been made. In slow chats, human moderators can accomplish this, while faster chats require moderation bots. This effect may be more important than the absolute amount of time between posting and ban.

Behavior	Response	% Increase after Event
Spam	Banned	13.1%
	Not Banned	46.7%
Question	Banned	34.0%
	Not Banned	55.3%
Smile	Banned	-10.0%
	Not Banned	220.0%

Table 9: Percentage Increase after Event of Same Type, Banned vs Not Banned

While this analysis shows significant generalized deterrence effects, these effects are not as large as imitation effects for spam or questions; even though a message was banned and thus disappeared from the screen, its brief presence reminded users that this behavior was one possibility in which they could engage, suggesting the presence of the conception effect [53].

Discussion

The analyses presented above show significant but not uniform deterrent effects from bans, and a significant decrease in spam behavior during chat moderation modes.

The deterrence effects of seeing another person being banned for engaging in a particular type of behavior are in line with literature on generalized deterrence [16][52]. While deterrence literature suggests that the effect of general deterrence decreases as social distance between observer and criminal increases, such distinctions are less relevant on Twitch. Users directly observe the punishment of users who are mostly socially indistinguishable from them, and they are not at all affected by bans in other networks (i.e. other channels), which they don't observe.

In this case, channel chat moderation modes had limited effects. While they were successful in deterring one type of behavior, spam, their lack of flexibility prevented them from being applied differentially to discourage or encourage other types of behavior; we did not find evidence that reducing the volume of spam made space for other more positive behaviors. Regular bans were more flexible than moderation modes in differentially discouraging different types of behavior, but none of the moderation tools were effective in encouraging positive behaviors (i.e., smiles).

ANALYSIS 4: DURATION OF IMPACT

In the final part of this study we examined how long imitation effects last. We calculated correlations between counts of each type of behavior in two blocks of ten consecutive messages with a varying number of messages between them, using these correlations across channels to measure imitation effects. We looked at decay in the effects by increasing the delay between the set of messages that were the source of the effect and the set of messages that displayed imitation. Table 10 shows correlations between counts of behavior in blocks of ten messages separated by zero, ten, twenty, thirty, forty, and fifty messages. For example, the correlation between count of spam messages in one block of ten messages and count of spam messages in the next block of ten messages was 0.29, but the correlation between count of spam messages in one block of ten messages and count of spam messages in another block of ten messages 50 messages later was 0.15. For all three behaviors, there was a nonzero correlation between behavior in the initial block and behavior in the next block, but this decreased as the next block moved further away. This shows that these behaviors have significant short-term effects but that these effects decrease steadily over time.

This matches the results discussed earlier both in Analysis 1 and in Bakshy et al. [8]. Users were influenced to post content that they saw other users post, and the more other users posted the content the more likely they were to follow suit. However, this effect was limited to a relatively brief period of time after the initial posting.

	Interval Size					
Behavior	**0**	**10**	**20**	**30**	**40**	**50**
Spam	0.29	0.23	0.20	0.17	0.17	0.15
Questions	0.13	0.11	0.09	0.08	0.07	0.07
Smiles	0.10	0.07	0.06	0.05	0.05	0.03

Table 10: Correlation between count of messages of each type in one group of ten messages and the count of messages of same type in the next group of ten at increasing intervals between groups

These correlations were calculated for each of the 400 channels in our sample and for each of the three behaviors described above: “spam”, “questions”, and “smiles”. Approximately 2 million comparisons were included in each of these analyses.

Taken together, these results suggest that a small cluster of messages of a particular type may have some short-term impact, but is unlikely to change the culture of the channel. Such a cultural shift might plausibly occur as a result of a larger number of clusters of particular types of behavior, especially if such behaviors were encouraged by users with authority or left unbanned.

CONCLUSIONS

The studies described above show clear patterns of imitation and deterrence in Twitch chats. When a user posted a message with a particular type of behavior, subsequent messages were substantially more likely to contain that behavior. This is consistent with much research on imitation.

Beyond this general imitation effect, our research shows that certain types of users were more influential on certain types of behaviors. Overall, users with more authority had more influence on most types of behavior. However, Twitch Turbo users had very little influence or even negative influence, which may reflect their outsider status.

Finally, users were responsive to deterrent measures. When a particular type of behavior was banned, the frequency of that behavior in the next group of messages decreased. When chat moderation modes were enabled, most types of behavior were unaffected by all three modes but spam decreased in all cases.

These findings have substantial practical implications. First, in designing future moderation tools, it may be useful to consider the possibility of encouraging desirable behaviors in combination with the standard method of punishing undesirable ones. While Analysis 4 shows minimal long-term impact of imitation effects on culture, it may be possible to curate a culture with repeated interventions over time. By actively banning undesired content and encouraging users to behave positively toward each other, site managers may be able to create a community that is resistant to the impact of undesirable behaviors from newcomers or outsiders who aim to disrupt. Conversely, once a channel has developed a norm of undesirable behaviors, it may be more difficult to stop these behaviors from spreading.

These findings also point to several areas for future research. Visible examples of good and bad behavior have substantial impact [30]. Future studies could experiment with different approaches to making good and bad behaviors more or less visible. Subsequent to the collection of the data used in this study, Twitch added a new suite of moderation tools that allowed moderators to review messages before they appear in chat and remove unwanted messages, precluding any possibility of imitation stemming from removed messages.

In addition, while the proactive tools explored here (i.e., the chat modes in Twitch) are relatively inflexible, alternative tools could be tested that warned users in advance if their message was likely to be banned based on its content. Moreover, new tools that offer broadcasters more flexibility in determining which types of behavior to discourage may be more effective in regulating culture. While there may be a role for machine learning-driven approaches to regulating specific types of behavior, such approaches are inherently risky in that they may drive users away because of unclear or inconsistent standards for appropriate behavior or may encourage some users to be increasingly creative in their attempts to be offensive in order to circumvent automated bans. A promising direction for future exploration is use of social pressure to enforce standards; allowing established users in the community more visibility and a more active role in discouraging unwanted behavior, even if not directly through bans, can turn a chatroom full of viewers into a group of allies.

As discussed above, one of the primary difficulties in establishing causality is elimination of outside explanations. Due to limitations of the IRC medium, this analysis does not control for behaviors exhibited through video or audio on the stream itself. As such, one alternative explanation for the imitation effects observed is that users were more likely to engage in certain behaviors when they received a cue from the streamer that such behaviors were acceptable, and that more established users were more sensitive to these cues. Broadcasters who are more calm and collected may recruit and retain viewers who are more likely to behave calmly [21]. While we cannot separate this influence out in our analysis, the implications of this explanation are mostly the same; in both cases, streamers can significantly affect chat behaviors both through encouraging examples of acceptable behavior and by using tools to deter unacceptable behavior, regardless of whether the reaction comes from a moderator in the chat or the streamer.

Our research suggests that existing moderation tools can be effective and, by extension, that moderators and broadcasters have some ability to shape the type of chat environment that they want, though they may still be vulnerable to persistent campaigns of targeted harassment. The combination of moderation tools described in this paper can help channels of almost any size; smaller channels will see more impact from bans because the next ten messages will last longer. Larger channels may benefit more from a combination of chat modes and bans.

More broadly, these findings have implications for how moderation should be explored both in research and in practice. Moderation can be viewed not only as a reaction to specific events but also a method for preventing the spread of unwanted behavior and development of undesirable norms for what conduct is acceptable. The development of behavioral standards through display of positive behaviors is possible both independent from and in combination with moderation tools.

ACKNOWLEDGEMENTS

This project was supported by the National Science Foundation under awards CNS-1221006 and VOSS-1322278.

REFERENCES

1. Akers, R. L., & Cochran, J. K. 1985. Adolescent marijuana use: A test of three theories of deviant behavior. *Deviant Behavior*. 6, 4: 323-346.
2. Akers, R. L., & La Greca, A. J. 1991. Alcohol use among the elderly: Social learning, community context, and life events. In *Society, culture, and drinking patterns reexamined*. 242-262.
3. Akers, R. L., & Lee, G. 1996. A longitudinal test of social learning theory: Adolescent smoking. *Journal of Drug Issues*. 26, 2: 317-343.
4. Akers, R. L., Skinner, W. F., Krohn, M. D., & Lauer, R. M. 1987. Recent trends in teenage tobacco use-findings from a 5-year longitudinal study. *Sociology and Social Research*. 71, 2: 110-114.
5. AnyKey. 2016. AnyKey Workshop #2 White Paper: Barriers to inclusion and retention: The role of community management and moderation.
6. Aral, S., & Walker, D. 2011. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*. 57, 9: 1623-1639.
7. Asch, S. E. 1956. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological monographs: General and applied*. 70, 9: 1.
8. Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. April. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, 519-528. ACM.
9. Bayer, J. B., Ellison, N. B., Schoenebeck, S. Y., & Falk, E. B. 2015. Sharing the small moments: ephemeral social interaction on Snapchat. *Information, Communication & Society*, 1-22.
10. boyd, d. 2014. *It's complicated: The social lives of networked teens.* Yale University Press.
11. Chartrand, T. L., & Bargh, J. A. 1999. The chameleon effect: The perception–behavior link and social interaction. *Journal of personality and social psychology*, 76, 6: 893.
12. Cialdini, R. B. 2009. *Influence: Science and practice* (Vol. 4). Pearson Education.
13. Citron, D. K. 2014. *Hate crimes in cyberspace*. Harvard University Press.
14. Coleman, G. 2014. Hacker, hoaxer, whistleblower, spy: The many faces of Anonymous. Verso Books.
15. Das, S., Kramer, A. D., Dabbish, L. A., & Hong, J. I. 2015. The Role of Social Influence In Security Feature Adoption. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15), 1416-1426.
16. Donath, J. 1996. Identity and deception in the virtual community. In *Communities in cyberspace*. 29-59.
17. Feigenbaum, J., Hendler, J. A., Jaggard, A. D., Weitzner, D. J., & Wright, R. N. 2011. Accountability and deterrence in online life. In *Proceedings of the 3rd International Web Science Conference*, 7.
18. Fotigames. 2015. "How to Handle Malicious Viewers/Trolls". From https://www.reddit.com/r/Twitch/comments/3p4px8/how_to_handle_malicious_viewers_trolls/. Retrieved 11 May 2016.
19. Fox, J., & Tang, W. Y. 2016. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media & Society*. March 8, 2016.
20. Goldenberg, D. 2015. "How Kappa Became the Face of Twitch". Five-thirty-eight. Retrieved 21 May 2016.
21. Hamilton, W. A., Garretson, O., & Kerne, A. 2014. Streaming on twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems,* 1315-1324. ACM.
22. Harry, D. 2014. "Past and Future of Game Spectating". Speakerdeck.com. Retrieved 21 May 2016.
23. Herring, S., Job-Sluder, K., Scheckler, R., & Barab, S. 2002. Searching for safety online: Managing "trolling"

in a feminist forum. *The Information Society*. 18, 5: 371-384.

24. Higgins, G. E., Fell, B. D., & Wilson, A. L. 2006. Digital piracy: Assessing the contributions of an integrated self-control theory and social learning theory using structural equation modeling. *Criminal Justice Studies*. 19, 1: 3-22.
25. Higgins, G. E., Wilson, A. L., & Fell, B. D. 2005. An application of deterrence theory to software piracy. *Journal of Criminal Justice and Popular Culture*. 12, 3: 166-184.
26. Hofling, C. K., Brotzman, E., Dalrymple, S., Graves, N., & Pierce, C. M. 1966. An experimental study in nurse-physician relationships. The Journal of nervous and mental disease. 143, 2: 171-180.
27. Hogg, M. A. 2001. A social identity theory of leadership. *Personality and social psychology review*. 5, 3: 184-200.
28. Houston, T. K., Cooper, L. A., & Ford, D. E. 2014. Internet support groups for depression: a 1-year prospective cohort study. *American Journal of Psychiatry*.
29. Kittur, A., Pendleton, B., & Kraut, R. E. 2009. Herding the cats: the influence of groups in coordinating peer production. In *Proceedings of the 5th international Symposium on Wikis and Open Collaboration* (p. 7). ACM.
30. Kraut, R. E., & Resnick, P. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press.
31. Levmore, S., & Nussbaum, M. C. 2010. *The Offensive Internet*. Harvard University Press.
32. Li, X., & Nergadze, N. 2009. Deterrence effect of four legal and extralegal factors on online copyright infringement. *Journal of Computer-Mediated Communication*. *14*, 2: 307-327.
33. Maiberg, E. 2014. “Twitch ranked 4th in peak internet traffic, ahead of Valve, Facebook, Hulu”. From http://www.gamespot.com/articles/twitch-ranked-4th-in-peak-internet-traffic-ahead-of-valve-facebook-hulu/1100-6417621/. Retrieved 11 May 2016.
34. Marwick, A. E. 2013. *Status update: Celebrity, publicity, and branding in the social media age*. Yale University Press.
35. Milgram, S. 1978. *Obedience to authority*. Harpercollins College.
36. Munroe, R. 2008. “ROBOT9000 and #xkcd-signal: Attacking Noise in Chat”. From blog.xkcd.com. Retrieved 11 May 2016.
37. Nagin, D. S. 1998. Criminal deterrence research at the outset of the twenty-first century. *Crime and justice*, 1-42.
38. Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. 2008. Normative social influence is underdetected. *Personality and social psychology bulletin*. 34, 7: 913-923.
39. Peters, D. P., & Ceci, S. J. 1982. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*. 5, 2: 187-195.
40. Phillips, D. P. 1974. The influence of suggestion on suicide: Substantive and theoretical implications of the Werther effect. *American Sociological Review*, 340-354.
41. Phillips, W. 2011. LOLing at tragedy: Facebook trolls, memorial pages and resistance to grief online. *First Monday*. 16*,* 12.
42. PsychCentral. 2008. Terms of Use, from http://psychcentral.com/about/terms.htm. Retrieved 8 Mar 2008.
43. Quantcast. 2016. “Twitch.tv Traffic and Demographics.” From https://www.quantcast.com/twitch.tv. Retrieved 11 May 2016.
44. Resnick, P. 2001. The social cost of cheap pseudonyms. *Journal of Economics & Management Strategy*. 10, 2: 173-199.
45. Romero, D. M., Meeder, B., & Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, 695-704.
46. Scott-Parker, B., Hyde, M. K., Watson, B., & King, M. J. 2013. Speeding by young novice drivers: What can personal characteristics and psychosocial theory add to our understanding?. *Accident Analysis & Prevention*. 50: 242-250.
47. Shadish, W. R., Cook, T. D., & Campbell, D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.
48. Sherif, M. 1936. *The psychology of social norms*. Harper.
49. Stafford, M. C., & Warr, M. 1993. A reconceptualization of general and specific deterrence. *Journal of research in crime and delinquency*, *30*(2), 123-135.
50. Suler. 2004. The online disinhibition effect. *Cyberpsychology & behavior*. *7*, 3: 321-326.
51. Taylor, T. L. 2009. *Play between worlds: Exploring online game culture*. MIT Press.

52. Tonry, M. 2008. Learning from the limitations of deterrence research. *Crime and Justice*. 37, 1: 279-311.

53. Wheeler, L. 1966. Toward a theory of behavioral contagion. *Psychological Review*. 73, 2: 179.

54. Ybarra, M. L., & Mitchell, K. J. 2008. How risky are social networking sites? A comparison of places online where youth sexual solicitation and harassment occurs. *Pediatrics*. 121, 2: e350-e357.

55. Zhu, H., Kraut, R., & Kittur, A. 2012. Organizing without formal organization: group identification, goal setting and social modeling in directing online production. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work,* 935-944. ACM.