Online deviant behavior on Social Media: The Macro, Miso, and Micro perspectives

By

QIUSI SUN
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Communication

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

____
Cuihua Shen, Chair

____
Martin Hilbert

____
Magdalena E. Wojcieszak

Committee in Charge

2022

Acknowledgement

This six-year rewarding journey is the most fulfilling experience in my life and it would not have been possible without who have continuously supported and encouraged me. I would like to express my sincere gratitude to all of them. I am deeply indebted to my academic advisor, Dr. Cuihua Shen, whose care and belief in me kept me staying optimistic and carrying on. She has guided and supported me with her wisdom, experience, and patience. Dr. Shen granted me the freedom to explore, at the same time, her critiques and supervision kept me on the right track. Outside of academia, she is also a lovely person to talk, seek advice, and share emotions and thoughts with. She has been such an amazing advisor that I could not undertake this journey without her support and encouragement.

I am extremely grateful to have Dr. Martin Hilbert and Dr Magdalena Wojcieszak on my committee. Dr. Hilbert has been like a second advisor to me, always providing me valuable insights when I was bogged down in the details. His suggestions kept me challenging myself and reevaluating my work from a bigger picture. Dr. Wojcieszak has always been willing to help and giving her best suggestions. I was so inspired by her perseverance and devotion to the research to stay strong and focus. Their constructive criticisms and valuable feedback at different stages of my dissertation were precious to me.

I am thankful to Dr. Jingwen Zhang, Dr. Zhou Yu, Dr. Wang Liao, Dr. Laramie Taylor, Dr. George Barnett, Dr. Richard Huskey, and Dr. Seth Frey, for all their academic advice. I would also like to thank my dearest friends and fellow cohort members, Dr. Jieyu Ding Featherstone and Dr. CJ Calabrese for treading through this hardship together. And thanks should also go to Qiankun Zhong for going through the pandemic together.

I am especially thankful to my family for their support, confidence in me, and unconditional love. My father, Wenzhe Sun, his determination and dedication to research and field work inspired me to pursue an academic career. Thanks to my mother, Yingqiu Fu, for being so supportive especially through difficult times. And special thanks to my two cats, Che-Che and Mimi, for always being there for me to cuddle.

I would be remiss in not mentioning our department of Communication at UC Davis. I received a lot of help from our knowledgeable professors, supportive colleagues, and friendly staff. I am extremely proud to be a graduate of the department, and I wish the department all the best for the future.

Abstract

Online deviant behaviors in social media have become one of the major concerns of the public, with its prevalence and rapid and widely spread. Though massive research has been done to understand the antecedents and consequences of online deviant behaviors, academic scholarship is still far from a definitive explanation about how the dynamic of the behavior is embedded in its context. In this dissertation, for two forms of online deviant behaviors, incivility and trolling, three studies were conducted to a) examine the overall dynamic of incivility on the macro level, b) understand the mechanisms of trolls and victims on the meso level, and c) identify the underlying mathematical architecture of trolling behavior on the micro level.

Study 1 provides a descriptive account of the incivility dynamic over the past eleven years by examining the trends of incivility in three major categories: political, non-political, and mixed. Using longitudinal data from Reddit that accounts for 95% of the entire Reddit universe across eleven years and relying on the combination of supervised machine learning models and traditional statistical inference, the study found that incivility consistently represent around 10% of total Reddit comments, with fluctuations that correspond to offline socio-political events and platform-specific policies. We also found that political groups tend to be more uncivil, and discussions in mixed groups that are not overtly political but nevertheless discuss politics are less uncivil than that in political groups.

Study 2 investigated the aftermath of trolling on community dynamics by examining the likelihood and conditions in which individual users react toward trolls. Using a longitudinal behavioral dataset collected from popular video communities on YouTube, the study found that the valence of the trolling message, characteristics of the individual member, as well as the patterns of past engagement with trolls from other community members all played a role in predicting how an individual would react to trolls. In other words, well-connected users situated in densely connected communities with a prior pattern of engaging trolls are more likely to respond to trolls, especially when the trolling messages convey negative sentiments.

Study 3 employed an information theoretical approach to deconstruct trolling behavior as a dynamic process to understand how people engage in such behavior. Using longitudinal data from Reddit’s active users and applied stochastic process and statistical inference, the study found that individuals engaging in trolling behaviors is a complex process that involves both internal influences and external influences from others they interact with. In addition, the hidden pattern of mathematical architecture shows a contagion effect of the behavior. Overall, the three studies demonstrated that online deviant behavior is nested in individual processing, community norms, and platform contexts.

Table of Contents

Chapter 1. Introduction

As social interactions in the digital environment are now prevalent as part of our daily activities, deviant behaviors in such places become increasingly visible. Online deviant behavior, taking in various forms and terms, such as online harassment, hate speech, trolling, incivility, and toxicity (Blackwell et al., 2017; Howard, 2019; Muddiman, 2017; Shen et al., 2020) is one of the central concerns of the academic and general public. According to Pew Research Center (2021b), more than 41% of American adults have experienced online harassment, including physical threats, sexual harassment, stalking, sustained harassment, name-calling, and purposefully embarrassment. Moreover, about 75% of those who have experienced said the most recent experience took place on social media. Another national poll on online and offline incivility found that about six in ten Americans experienced uncivil behaviors online, and 57% of them attributed incivility to social media (*Civility in America 2019: Solutions For Tomorrow*, 2019). More than half of the participants consider online deviant behaviors on social media as a major problem in both surveys. These alarming statistics urge us to systematically understand the dynamic of the deviant behaviors on social media.

This dissertation focuses on two categories of online deviant behaviors, namely, incivility and trolling, addressing the crucial problem from macro, meso, and micro perspectives by looking at the overall dynamics of incivility as well as the mechanisms between trolls and victims. By using longitudinal data from several social media platforms and a combination of computational methods and traditional statistical inference, the dissertation provided systematic evidence to unravel the myths of the severe problem of incivility and shed light on trolls' behavior patterns and intervention strategies.

1.1. Defining Online Deviant Behaviors

Though online deviant behaviors have drawn a great amount of research attention (Ortiz, 2020; Rossini, 2020), well-established definitions has not emerged yet. Despite all well-intentioned effort on a universal definition, research has showed that online deviant behaviors are a complex concept with multiple dimensions and subcategories without clear boundaries (Udris, 2017). The word *deviant* has already implied a comparison with *normal*. Without a standard of *normal*, deviant behaviors cannot be defined or identified. Rather, it is contextually sensitive and individual characteristics mattered (Masullo Chen et al., 2019). The perceptions of deviant behaviors varied across people, context, and culture. For instance, attacking one's appearance is normal in r/roastme sub-Reddit, while will be considered deviant in other groups. In addition, people who are older (Klempka & Stimson, 2014) and scored higher on Big Five Personality Traits of Agreeableness (Kenski et al., 2020), are more likely to have a higher standard for civility, resulting in strong reactions to behaviors that could be considered deviant.

On the other hand, because of the difficulty in defining the concept, scholars are reluctant to clump deviant behaviors together. Instead, several terms and concepts were used to refer to specific behaviors that could be considered deviant (Alonzo & Aiken, 2004; Blackwell et al., 2017; Chun et al., 2020; Howard, 2019; Muddiman, 2017; Shen et al., 2020), such as trolling, flaming, online harassment, cyberbullying, incivility, hate speech, etc. However, the concepts are ambiguous and inconsistent in different domains, leading to widely varying terms. For instance, "trolls" can be defined as people who disrupt online discussion and provoke aggressive responses in online forums (Engelin & De Silva, 2016). While in other contexts, the term is used to refer to ideological devices that express extreme political views (Fichman & Sanfilippo, 2016). In addition, the vogue definitions also lead to concept overlapping. As a result, different terms were used to refer to the same or similar behaviors, plaguing scholarship and leading to further problems in detection and categorization of online deviant behaviors. In this dissertation, we will focus in two specific concepts, namely, trolling and incivility.

The concept of “trolling” originated from folk culture, analogizing the imaginative creature who will attack from unexpected places to refer to unpleasant behaviors and people on the Internet after the 1980s (Donath, 2002). While unsurprisingly, in previous research, what constitutes trolling behaviors has been defined differently, and the definitions focused on two aspects. One stream focuses on the deceptive aspect of trolling behavior. An early study on trolls in USENET newsgroups from Donath (2002) defined them as troublemakers who pretend to be legitimate participants, provoke others by providing incorrect advice and damage the feeling of trust. Engelin and De Silva (2016) looked at massive Twitter accounts, identifying trolls as people who “interrupt, harass, or try to impose opinions to others” (p. 4) by using fake personas on Twitter. In other cases, trolls were considered hackers who used deceptive comments to find valuable information to attack victims on Reddit (Breeze, 2012).This stream of thought indicates that deception is one of the key characteristics of trolls, using misinformation, logical fallacies, or misleading information to manipulate individuals’ opinions.

Another school of thought believes that trolling should be defined by its provocative nature. Behaviors that “causing disruption” to communities and “triggering or exacerbating conflicts”(Hardaker, 2010, p. 237), intensifying engagement for personal amusement (Nevin, 2015), transgressing community norms that result in anger, harm, or discomfort (Bergstrom, 2011), and posting erroneous or inflammatory information to provoke a strong reaction (Merritt, 2012) were defined as trolling in different studies of misconducts on social media. Those definitions emphasized that trolling behavior could mislead individuals into time-wasting discussions, fail problem-solving, and negatively affect their experience.

A comprehensive definition was proposed to combine deception and provocation. An early qualitative study on trolls in a feminist forum defined trolls as people who lure others into pointless and time-consuming discussions by insulting, provoking, or rebuking (Herring et al., 2002). Trolling behaviors can be outright swearing, personal attacks, veiled insults, sarcasm, and off-topic statements(Cheng et al., 2017). And (Hardaker, 2010) built a model of "deception, aggression, disruption, and success" to define trolls, where trolling is "negatively marked online behavior" that successfully dysfunction discussions and lead to unpleasant experiences.

The definition of incivility also varies within and across disciplines, contexts, and studies (Cortina et al., 2001; Kim et al., 2021). First, incivility is distinct from impoliteness, negativity, and intolerance. Unlike impoliteness, which focuses on individual manners, incivility focuses on "a collective founded on democratic norms and mandates" (Papacharissi, 2004, p. 271). A distinction between incivility and negativity can be drawn, with the former being "detrimental to deliberative debate, attacking one's character" (Gervais, 2017) while the latter can be either civilly or uncivilly delivered to target an individual or an issue. In addition, a recent study also differentiated uncivil and intolerant speech in political talks (Rossini, 2020), with the former referring to discourse that goes against accepted social norms and the latter being discourse that promotes discrimination, derogation, and violence. Distinctive from impoliteness, negativity, and intolerance, incivility refers to behaviors (or more likely, comments and posts) that go against social norms to attack others.

Arguments for the definition of incivility are in line with some theoretical literature on the idea of inclusion, mutual respect, and sensitivity to inequality in democratic deliberation (Ferree et al., 2002). For instance, research on newspaper website comments from Coe and colleagues (2014) define incivility as "Features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants or its topics" (Coe et al., 2014, p. 660). It is not given individuals' opinions. Instead, it is a characteristic of interaction(Mutz, 2015). Scholars also point out that uncivil claims are usually not based on evidence or true information (Muddiman, 2017). Furthermore, in computer-mediated discourse, empirical studies about incivility consider it conceptualized as a continuum that ranges from mild forms to severe forms, from disrespect or frustration to strong derogatory remarks and foul language(Sobieraj & Berry, 2011).

While "incivility" seems inclusive and some researchers consider trolling uncivil behavior (Sadeque et al., 2019), the key difference between incivility and trolling is the desired outcome. From the original, trolling has a designated outcome as attracting attention and provoking responses, but it is not necessarily uncivil. Strategic acts such as sarcasm can also serve the purpose of trolling. However, uncivil discourse, in which people use foul language and harsh tone, can be considered a rhetorical act that is discrepant from social norms. It can serve as a strategy to troll or fulfill other goals in discussions. As it is approached as a function of the tone and the discourse features, there is no presupposed outcome of incivility.

1.2. Antecedents of Online Deviant Behaviors

Theoretically, the antecedents of online deviant behaviors in a computer-mediated environment were studied from different perspectives, such as the construction of social identity, social presence in online interaction, disinhibition effects, or deindividuation effects. Specifically, Research has demonstrated that characteristics of CMC, such as anonymity and invisibility, have a great influence on individuals' online behaviors (Febriana, 2019; Rains et al., 2017).

Anonymity is a combination of technical anonymity and social anonymity (Christopherson, 2007). The former is when there is no identifying information, and the latter is the perception that one is anonymous to others. The reduced cues model assumes that without a person's identity, people are apt to behave and communicate in ways that are otherwise different from the way they would present themselves if their identity were intact (Santana, 2014). Proponents of this model argue that anonymity in CMC is associated with a series of behavioral outcomes distinguished from face-to-face communication. Those outcomes include a tendency for individuals who are anonymous to be less inhibited in their expressions and adapt a deregulating way to communicate. Furthermore, anonymity causes people to feel unaccountable for their negative actions, as they cannot be identified to take responsibility for those actions. In addition to reduced cues, another important characteristic of online anonymity is that people can play-act at being someone else or put on different online personae that are a distinct identity from their "real life"(Zhao et al., 2008). Anonymity allows people to recreate their biography and personality. The outcome of this "role-playing" can also be different from people's face-to-face communication. In general, anonymity can foster a sense of impunity, loss of self-awareness, and a likelihood of acting upon normally inhibited impulses that may be inconsistent with their offline self.

Invisibility is different from anonymity, whereas it inhibits social presence and diminishes the salience of others in the interaction and the salience of the interpersonal relationship. It shows the inability of the communication channel to transmit contextual and nonverbal cues (Croes et al., 2016). The low social presence can reduce the anxiety caused by being seen when typing messages or responding to others. In addition, it renders the possible stereotype or prejudice that is related to demographic attributes (i.e., gender, race, age, (McKenna & Green, 2002), physical characteristics (i.e., appearance, weight, height), and stigmatized behaviors (McKenna & Seidman, 2005). However, the diminished social presence can lead to a state of decreased self-evaluation and depersonalization in which both others and selves are perceived as part of a broader social group that is salient during interaction (Lapidot-Lefler & Barak, 2012).

What Builds on the idea of anonymity and invisibility is the framework of the online disinhibition effect (Suler, 2004).The online disinhibition effect serves as a theoretical guide lens positing that people tend to lose their inhibitions during interactions in an online environment. In other words, people tend to behave in an online setting that they would not otherwise behave in face-to-face communication as there is no or little concern about self-presentation or judgment by others. The effect may work in two directions, benign and toxic. Benign disinhibition happens when people self-disclose intimate feelings, emotions, or fears or engage in unusual behaviors that show support, kindness, and generosity. While toxic disinhibition focuses on the negative effect, such as rude or harsh language, hostility, and even threats, which was described as "simply a blind catharsis, a fruitless repetition compulsion, and an acting out of unsavory needs without any personal growth at all (Suler, 2004, p. 321)". The theoretical framework proposes that one of the key reasons people behave deviated online is physical, visual, and emotional distances that reduce empathy with others. It suggests that people engaging in deviant behavior online is a result of a person shifting to dissociation from inhibiting guilt, anxiety, and other moral senses.

The online disinhibition effect is not the only theoretical framework that proposes a theoretical mechanism for online deviant behaviors. A competing theoretical model, the social identity model of deindividuation effect (SIDE), provides different arguments and explanations. The SIDE model was developed to account for a variety of effects of technology mediation on social interactions and how CMC moderates the effect of social identities on interactions (Postmes et al., 1998). In many CMC contexts, especially those that are text-driven, the feeling of anonymity and invisibility can reduce the awareness of personal identity but increase the presence of social identity. According to Lea and Spears (1991), people in anonymous and invisible online environments are particularly influenced by cues related to group membership. Individuals are more likely to operate at the level of social identity rather than their personal identity and are more likely to conform to group norms. Strategically, people will behave in a way to support their group identity when they are anonymous and invisible in CMC (Postmes et al., 1998). People may intentionally express their social identity to seek acceptance from ingroup members and enhance the standing of their ingroup. Those identity performances include communication behaviors such as conforming to group norms, expressing prototypical group opinions, rendering oneself identifiable as a group member, explicitly invoking the group, resistance against the outgroup, and outgroup denigration. In other words, the collective mind of a group takes possession of the individuals. As the SIDE model suggested, individuals' social identity becomes more salient that deviant behaviors can be conducted as consolidation of group identity as well as mobilization of group identity. For instance, insulting an outgroup can reinforce one's status as an ingroup member and positively distinguish one's group from an outgroup. The SIDE model helps to explain how ingroup favoritism can result in deviant behaviors and how deviant behavior can serve as a method for identity consolidation and mobilization in an online communication environment.

1.3. Previous Studies on the Effect of Online Deviant Behaviors

The literature on online deviant behaviors has grown tremendously since the development of web 2.0 as such phenomenon gained a lot of publicity. Thereinto, a large body of previous studies have been conducted on the consequences in different domains and from different perspectives. For example, a content analysis of tweets about the 2013 Colorado flood event (Anderson & Huntington, 2017) suggested that incivility was used in association with skeptical perspectives of political topics (e.g., climate change) and by those who are more right-leaning. In addition, other deviant behaviors, like trolling and toxicity, have been studied in domains such as online gaming and online communities.

The political outcomes attributed to online deviant behaviors have yielded a substantial amount of evidence on political partisanship and polarization, trust in news sources, and emotions and cognitions. However, the findings are inconsistent. For example, Hwang et al. (2014) found exposure to uncivil online comments does not affect attitude polarization but affects the perceived polarization of the public. Whereas research on partisan media and incivility suggests that (de)polarization is a complex effect involving one's partisan affiliations, conflict orientation, and the target of incivility (Druckman et al., 2019). It shows that out-party incivility will reduce the likeability and trustworthiness of the out-party and increase affective polarization, while in-party incivility will have the opposite effect.

Furthermore, the credibility of the news source, political trust, and political efficacy were found to be associated with uncivil blogger commentary (Borah, 2013). An online experiment conducted by Anderson et al. (2018) looked at the association between incivility and media perception indicated that people exposed to uncivil comments are more likely to perceive the news source as biased, especially for those who hold a strong conservative ideological viewpoint. Interestingly, incivility in blog commentary increases the credibility of a news story (Borah, 2013; Thorson, 2016; Thorson et al., 2010), but the blogger is viewed less favorably and less trustworthy by the readers (Borah, 2013; Graf et al., 2017). In addition, the sense of untrustworthiness has been shown to be able to be generalized to politicians and politics (Van't Riet & Van Stekelenburg, 2021). Finally, exposure to uncivil online comments decreases the expectations about public deliberation.

Exposure to incivility is also shown to have negative psychological effects. Chen and Ng (2017) found uncivil disagreement has a greater impact on negative emotions than civil disagreement comments. Similar findings from Gervais (2017) states that exposure to disagreeable uncivil political talk induces feelings of anger and aversion, reducing satisfaction with the political discourses. On the other hand, multiple studies have shown that exposure to uncivil disagreement increases the likelihood of future incivility use. Exposure to mainstreaming uncivil media has been shown to increase the likelihood of engaging in verbal aggression in online discussions (Cicchirillo et al., 2015). Later research conducted by Rösner et al. (2016) showed that exposure to incivility could increase individuals' hostile cognitions.

In the online gaming environment, more and more studies pointed out that deviant behaviors were treated as part of the geek culture (Beres et al., 2021). Descriptive research (Ballard & Welch, 2017) reported that more than half of the participants had been victimized by cyberbullying, and 35% had engaged in cyberbullying during online gaming. Female players, players from LGBT groups (Ballard & Welch, 2017), and players with lower rank and less experience (Kwak et al., 2015) will have a higher victimization rate. In addition, opponents are more like to be the targets of toxicity than teammates (Ballard & Welch, 2017), while teammates are more likely to be affected by toxic behavior (de Mesquita Neto & Becker, 2018). With the competitive nature of online gaming, toxic behaviors are shown to be contagious (Shen et al., 2020) and have become an organic component within the game culture (Kou, 2020).

As toxic behavior is a dynamic social process during game playing, it also largely impacts individual gamers' behaviors and psychological well-being (Shores et al., 2014). For instance, behavioral data suggested that players are not motivated to report toxic behavior (Kwak et al., 2015). Compared to teammates exposed to intergroup toxicity, opponents are more likely to report toxicity experienced from outgroups, and players from losing teams are more likely to report toxicity (Kwak et al., 2015). Previous research showed that toxic players could drive away new players, but more experienced players are more resilient to toxic behaviors (Shores et al., 2014). On-going research also suggests that, in general, experiencing toxicity in games will discourage game participation and reduce the psychological benefit of game-playing (Sun, 2022). Pervasive toxic behavior in massively multiplayer online games (MMOs) can reduce players' enjoyment, satisfaction, and commitment to the game (Shores et al., 2014).

Although major research about online deviant behaviors focused on behavior identification, the effects of the behaviors in online communities have also been studied. Ansong et al. (2013) looked at trolls in Ghana, noticing that trolling behaviors caused some legitimate community members to exit. Research conducted on the political discussion in USENET shows that trolls are isolated and relegated to the network periphery (Kelly et al., 2006), while a study on discussion network of a fan group found the opposite reaction, where members of an online fans group tried to engage with a troll to understand and change the troll's behavior rather than ignoring it (Baker, 2001). Systematic review and interviews of trolling identification (Cruz et al., 2018) show that it can be conceived as a constellation of learning, assimilating, and transgressing, which is confirmed by an online experiment conducted by Cheng et al. (2017) that negative mood and negative context increase the likelihood of trolling, and exposure to trolling behaviors is likely to trigger similar trolling behaviors.

Moreover, extensive research has been dedicated to finding out ways to prevent or intervene in online deviant behaviors. For instance, Kim and Gonzales (2022) and suggested on platform level bystander-centric strategies such as aggregated feedback systems (like/dislike system, ranking system, etc.) that provide reactions from others to uncivil comments that decrease individuals' intention to engage in arguing. On the community level, policies, and practices (user registration, moderating rules, etc.) are suggested to be able to reduce incivility (Ksiazek, 2015). However, Masullo Chen et al. (2019) retorted that deviant behaviors should be permitted in digital space considering such behaviors also had beneficial social purposes (draw attention, build bonds, etc.) with various contextual definitions.

1.4. Limitations of Extent Literature

While the extant literature has offered valuable insights on online deviant behaviors, at least three major limitations need scholarly attention. The first notable limitation with previous studies is that most of them have focused primarily on characteristics of deviant behaviors, while little attention has been given to the communication process involving such behaviors. In particular, people could get affected by prior online behaviors that both themselves conducted deviant behaviors in the past and exposed to previous deviant behaviors from others raise the likelihood of future deviant behaviors. Such behaviors can persist across those affected people to spread further. Related studies investigated whether an individual can be influenced to become a troll (Cheng et al., 2017) in an online community and whether toxic behavior can be contagion in MMOs (Shen et al., 2020). Cheng et al. (2017)'s study looked at the overtime proportion of trolling messages finding that trolling behavior can be spread in an online community, and Shen Shen et al. (2020)'s study, more specifically, found that exposure to toxic behaviors can lead to an adoption of such behaviors. However, both studies are taken from a massive perspective that individuals were treated as members of a large collective.

The second limitation is that the majority of previous research focused on perpetrators and individual level, while the victims are not passively being bullied, and their reactions need also be studied. Individuals may take different actions and even work as a collective unit to react to deviant behaviors, so the detrimental impact of deviant behaviors on individuals and higher-level communities are varied across audiences and contexts. For example, limited research on victims' reactions suggested that victims are likely to leave the discussion (Ansong et al., 2013), argue with perpetrators (Baker, 2001), and ignore the deviant messages (Kelly et al., 2006). But little do we know the mechanism leading to different reactions, nor the collective effect of individual victims' reactions. As a result, we are missing the big picture that can help identify the dynamics involving both perpetrators and victims at large.

Third, even though the general public observation of online deviant behaviors resulted in an impression of an increasing volume of deviant behaviors online, little research has been done to systematically understand the distribution of deviant behaviors in online space, with some exceptions like Siegel et al. (2018)'s and Theocharis et al. (2020) work. Meanwhile, the limited research on the temporal dynamics of deviant behaviors mostly focused on a certain contentious time and resulted in mixed evidence. For instance, three studies about incivility on social media platforms during the 2016 US presidential election and its aftermath resulted in different conclusions. Siegel et al. (2018) found no constant increases in incivility on Twitter during the period; rather, the spikes of incivility are random and unrelated to political events. In contrast, Theocharis et al. (2020) Twitter data suggested that the proportions of incivility are stable and spikes of incivility correspond to political events. In addition, research on Reddit (Nithyanand et al., 2017b) found that the presidential campaign fueled the use of incivility.

1.5. Overview of the Dissertation

In the following chapters, this dissertation attempted to address these limitations and is organized as follows. First, to understand the pervasiveness of deviant behavior in the online environment, Chapter 2 provides a systematic panorama of online deviant behaviors on a social media platform in a decade. Relying on Natural Language Processing (NLP) models and traditional statistical inference, the chapter focuses on the descriptive account of online incivility 1) overtime, 2) across different contexts, 3) as triggered by external events. The results show that the proportion of incivility remained constant across the years, with variations among different contexts, and the fluctuations are affected by platform-level policies and external events.

Chapter 3 presents the study focusing on victims of online deviant behaviors. It applies a network perspective to investigate how individuals in an online community react to trolls as a way to defend their social identity and provide insights into the dynamic of trolling in online communities. The results suggest that the network structural position can provide information about community dynamics and individuals' involvement that can predict how individuals and the community as a whole react to trolls.

Chapter 4 aimed to investigate whether and how online deviant behaviors are self-activated or situational and explore the predictive model of the individual trolling behavioral dynamic. The architecture of predictive models of individual behaviors shows that the knowledge of the recent past of both individuals' own and behaviors of people they encountered provides sufficient information for predicting individual future trolling behavior. The result suggested trolling is a contagion behavior that can be picked up by ordinary individuals. The final chapter, Chapter 5, summarizes the major contributions of this dissertation, discusses the limitations, and suggests future directions for research.

Chapter 2. Over-time Trends in Incivility on Social Media: Evidence from Political, Non-political, and Mixed Sub-reddits Over Eleven Years

Scholars and observers worry that public debates in the United States are growing increasingly uncivil. Politicians attack their opponents, partisans report unprecedented hostility toward opposition-party supporters (Iyengar et al., 2019), and partisan media describe the opposing out-party as Nazis or Communists (Berry & Sobieraj, 2013) and feature "in your face" debates (Mutz & Reeves, 2005). Concerns with incivility often pertain to the Internet. Incivility is pervasive in online communities (Reader, 2012). In 2018, 84% of Americans reported having experienced incivility online, and those who did encounter it roughly 11 times a week (KRC Research, 2018). These encounters can have negative effects. For example, the use of and exposure to incivility generates anger, anxiety, or mental distress, and can lead to aggression (Gervais, 2015) and hostile communication (Groshek & Cutino, 2016). In addition, incivility can drive users away from online discussions and lead to general dissatisfaction with public discourse (Anderson et al., 2014; Bauman et al., 2013; Moor et al., 2010; Ransbotham et al., 2016).

In this project, we aim to address a fundamental descriptive question regarding over-time variations in incivility across a range of online communities. We rely on the most comprehensive longitudinal dataset of Reddit comments from 2008 to 2019[1] and a combination of computational methods (i.e., a neural, BERT-based classifier to capture incivility in an incredibly large corpus of data, see Davidson et al., 2020) and traditional statistical inference (e.g., ANOVA and student t-test) to provide a descriptive account of online incivility (1) over-time, (2) across different

[1] We note that we have 13 years of Reddit data (i.e., 2006-2019), yet only posts and comments starting in 2008 can be analyzed for our purposes. This is because there were only 2 sub-reddits in 2006 and 4 in 2007, all created by administrators. Given that the first political sub-reddit (politics) was created in August 2007, the comparison among sub-reddit categories was done from 2008.

contexts of online discussions (i.e., political, mixed, and non-political), and (3) as influenced by external events.

Our extensive data show that the volume of incivility increased with the overall increase in the volume of online exchanges, but its proportion remained rather constant across the years, oscillating at roughly 10%. Consistent with the general observations, discussions about politics generate consistently *more* incivility than non-political and mixed discussions. That said, when aggregated across the years, incivility in gaming communities that sometimes discuss politics is significantly *higher* than in other groups, even explicitly political ones. Supporting worries about the difficulty of cross-party exchanges, politically heterogeneous online communities – where liberals and conservatives meet – generate more incivility than politically homogeneous liberal or conservative communities. Moreover, fluctuations in incivility are affected by platform-level policies and external events.

2.1. Incivility on Social Media

There is some conceptual and operational ambiguity in the existing literature on incivility and related concepts under the umbrella of toxic, offensive, or intolerant speech (see Kim et al., 2021 and Rossini, 2020 for recent reviews). Sometimes incivility is used to refer to impolite or negative speech. Yet, unlike impoliteness, incivility is seen as "individual behaviors that threaten a collective founded on democratic norms" (Papacharissi, 2004, p. 271), and - unlike negativity - incivility is said to be detrimental to deliberative debate and reduce the deliberative potential of offline or online conversations (Gervais, 2017). Recent work differentiates between uncivil and intolerant speech, with the former including discourse that goes against accepted social norms and the latter being discourse that promotes discrimination, derogation, and violence (Rossini, 2020).

Here, we do not address the distinction between incivility and other related concepts, nor do we test its democratic effects. We follow Coe et al. (2014), seeing incivility as “features of discussion that convey disrespectful tone toward the discussion forum, its participants, or its topics" (Coe et al., 2014, p. 660). Accordingly, we adapt the operational definition of incivility as speech that includes “name-calling, mean-spirited or disparaging words directed at a person or a group of people, an idea, plan, policy, or behavior; using vulgarity, profanity, improper language and pejorative remarks about the way a person communicates" (Coe et al., 2014, p. 660). As such, our project includes an empirically captured speech that merely counters social norms, e.g., name-calling, as well as the arguably more problematic intolerant speech that can be hateful toward social groups.

2.1.1. Macro Trends

Many observers lament declines in the quality of public discourse in the United States (Anderson et al., 2014; Santana, 2014) and some scholars are concerned that the affordances of social media platforms, such as anonymous or pseudonymous communication, have led to increases in incivility and its normalization in the online public sphere (Leurs & Zimmer, 2017; Theocharis et al., 2020). And yet, systematic evidence of these potential increases in incivility on social media platforms is still limited. Research on the temporal dynamics of incivility mostly focuses on Twitter - a platform used by a minority of American adults (22%; Pew Research Center, 2021) - and typically during certain contentious times and/or salient political events. The resulting evidence is mixed. For instance, Siegel et al. (2018) find no constant increases in incivility on Twitter during the 2016 presidential election and its aftermath; rather, their data suggest random spikes in incivility unrelated to external events. In contrast, analyzing longitudinal data from Twitter after the 2016 presidential election, Theocharis et al. (2020) show that the prevalence of uncivil tweets mentioning Members of the US Congress is rather stable, and spikes in incivility correspond to political events (e.g., a white nationalist rally) and policy debates (e.g., healthcare). Yet in other work looking at Reddit, Nithyanand et al. (2017b) find a sharp increase in incivility in political sub-reddits during the 2016 presidential campaign. Although those studies provide important insights into the dynamics of incivility on social media platforms, the timeframes analyzed are rather short and it is not clear whether extant worries regarding growing incivility and its normalization are warranted. By examining a much longer time span of nearly the universe of online expressions on Reddit, one of the most popular social media platforms, this project offers a macro-level panorama of variations in online expressions of incivility. We first ask: *RQ2.1: Has there been an increase in incivility, in the aggregate, on Reddit between 2008 and 2019?*

2.1.2. Contextual Influences

In addition to offering systematic evidence on whether, and the extent to which, incivility increased on social media over the past 11 years, our major contribution lies in testing these variations across different kinds of groups. Different topics and community cultures in online groups, which are developed by niche interests and user engagement, may promote or discourage uncivil behavior (Massanari, 2017). With different discourse dynamics, it is possible that the variations in incivility differ across various types of discussions, political and non-political alike. Our project is, to our knowledge, the first to differentiate expressions of incivility in political versus non-political groups, and, furthermore, across various categories of political groups (e.g., liberal, conservative, or heterogeneous), groups focusing on non-political issues (e.g., fashion, gaming), and also groups where users touch on both (e.g., discussing global warming in sub-reddits dedicated to cars; see Wojcieszak & Mutz, 2009).

2.1.2.1. Political, Mixed, and Non-political Discussions. Extant concerns with, and past work on, incivility mostly focuses on *political* incivility. This work finds substantial amounts of incivility in the comment sections of news websites (i.e., around 20% of comments were found to be uncivil in online newspaper comment sections, Coe et al., 2014) and on social media platforms (e.g., around 9% in political comments on Reddit, Nithyanand et al., 2017a, between 15% to 20% on Twitter, Theocharis et al., 2020). Yet, the focus on incivility in political spaces is rather narrow given that many Americans see politics as complex, boring, or overly divisive (Klar, 2018; Klar & Krupnikov, 2016), and avoid information about news and politics altogether (Feldman et al., 2013; Guess, 2021; Prior, 2007; Wojcieszak et al., 2021). Accordingly, most users do not discuss politics online (Barberá et al., 2019) and do not follow any political accounts on social media (Eady et al., 2019; Thorson, 2016). Clearly, examining strictly political incivility or incivility in overtly political spaces misses a large part of the online information and communication ecosystem.

For one, the nature of the online discussion is never clear-cut and people do engage in political exchanges in groups organized around *non-political* topics (Wojcieszak & Mutz, 2009). There, users connect with others based on shared non-political interests (e.g., following the same celebrity or being parents) and yet encounter politics inadvertently (e.g., when a celebrity endorses a politician on their Facebook page or a parenting sub-reddit discusses funding for education) (Fletcher & Nielsen, 2018; Silver & Andrey, 2019; Wojcieszak & Mutz, 2009). We refer to these groups as *mixed*, those where politics is *not* the central purpose but where users nevertheless engage in political talk. Even though users report encountering disagreement when political discussions emerge in non-political spaces (Wojcieszak & Mutz, 2009), the research found these mixed groups generated less incivility than explicitly political discussions

(Rajadesingan et al., 2021). After all, once people establish a shared interest, they may be more open to potential disagreements when politics emerge and engage with others more politely and with an open mind.

Second, as aforementioned, most people do not go online to exchange political information and may also shy away from discussions that entail any political topics altogether. Accordingly, the most popular online groups on social media platforms pertaining to entertainment. For instance, the most followed Facebook pages are Facebook App, Samsung, and Cristiano Ronaldo, focusing on topics such as games, technology, and celebrities ("List of most followed Facebook Pages", 2021). Similarly, among the top ten most followed Twitter accounts, eight are celebrities, and only one (Barack Obama) is a political figure ("List of most followed Twitter accounts", 2021). The same pattern is found in YouTube and Reddit, with all top 10 most subscribed YouTube channels and eight sub-reddits being entertainment (Baer , 2021; "List of most subscribed YouTube channels", 2021). Given the popularity of non-political spaces, we attend to these largely overlooked discussions. Even though there may be important topical differences between non-political groups, as we detail below, on average these groups may not entail as much name-calling, personal attacks, or disparaging or mean-spirited language as the political or even the mixed communities. One could expect the members of groups focused on movies, celebrities, pets, or technology to be bonded by common fandom (Seregina & Schouten, 2017) and *a priori* more favorable toward one another due to shared interests.

In sum, although mixed and non-political discussions may be less uncivil than political ones, this idea remains untested. Similarly, it is not clear whether fluctuations in incivility would differ across political, mixed, and non-political groups. If - as some fear - online discourse is increasingly uncivil, we would see growth in incivility across these three types of groups. If, however, the shared interests and common ground matter to online discourse, the trends would be less pronounced in mixed and especially in non-political groups. Given the lack of clear-cut directional expectations and the largely descriptive nature of our work, we ask: *RQ2.2: Have there been changes in incivility between political, non-political, and mixed groups?*

2.1.2.2. Specific Types of Online Discussions.

2.1.2.2.1. Ideologically homogeneous vs heterogeneous political and mixed groups. To portray the tested dynamics comprehensively, we offer a nuanced differentiation within political and mixed as well as non-political groups. First, we distinguish between political and mixed groups that are ideologically homogeneous versus heterogeneous. Considering the current polarized climate in the US, discussions between people who hold different views may be substantially more uncivil than discussions between people with similar political affiliations (in that Democrats may clash with Republicans and liberals may call conservatives names). That said, ideologically homogeneous groups could also entail high levels of incivility (in that Democrats/liberals could unite against former President Trump or Republicans/conservatives could bash the policies of President Biden, for instance). Research suggests that ideologically homogeneous networks may cultivate beliefs in conspiracy theories or foster extremist attitudes (Warner & Neville-Shepard, 2014); these beliefs and attitudes may result in strong and emotional opinion expression, which, in turn, could lead to incivility (Stevens et al., 2021). In short, whether ideologically homogeneous or heterogeneous political and mixed discussions are more uncivil is not only unexamined but also unclear. Our next question, therefore, asks: RQ2.3: Have there been changes in incivility between ideologically homogeneous and heterogeneous groups?

2.1.2.2.2. Conservative vs Liberal Homogeneous Groups. Within i*deologically homogeneous* political and mixed groups, we attend to expressions of incivility in liberal and conservative groups. Previous studies on group identity and norms show that conservatives and liberals follow different social norms for incivility (Rains et al., 2017) and see incivility differently; for instance, conservatives are less likely to perceive messages as uncivil (Kenski et al., 2020). In addition, Donald Trump's presidency may have encouraged or normalized incivility among conservatives (e.g., during Trump's election, there was more incivility in conservative sub-reddits; Nithyanand et al., 2017b). Thus, conservatives may be more likely to express incivility as they may see it as a usual or more accepted way of expression than liberals. On the other hand, several studies showed that on social media platforms liberals were more likely to "like" or "thumb-up" uncivil comments (Kim et al., 2021; Rains et al., 2017), indicating liberals agree with or endorse uncivil expressions; this may lead liberals to express uncivilly to gain agreement from their peers. And yet, a study on unacceptable and uncivil behavior in US politics finds that Republicans and Democrats react in similar ways to uncivil messages (Muddiman, 2021). We, therefore, ask: *RQ2.4: Have there been changes in incivility between politically liberal and conservative groups?*

2.1.2.2.3. Different Non-political Topics. Lastly, we examine whether incivility levels differ across various topics within the mixed and non-political groups, testing discussions revolving around entertainment, sports, lifestyle, and technology, among others (as detailed below). Some of these topics may touch on individual identity, in a way similar to that political stances can (e.g., sports or gaming, Vale & Fernandes, 2018; Murphy, 2004) and thus generate heated discussions that may lead to uncivil discourse. Inasmuch as, say, fans of the Dallas Cowboys see the Philadelphia Eagles as a rival, discussions about sports could be more uncivil than those about politics. Furthermore, certain hobby communities have "geek" cultures where incivility may be the norm (Massanari, 2017). For instance, participants in gaming communities may bash others for losing or call them names for poor performance (Shen et al., 2020). In short, non-political groups discussing distinct topics might differ in the volume and fluctuations in incivility. Also, some of those communities (e.g., sports, gaming) may be similar to explicitly political groups. These questions remain unaddressed in extant work. *RQ2.5: Have there been changes in incivility between different non-political groups?*

2.1.3. External Events

In testing these questions, we attend to the extent to which external events may influence the prevalence of and changes in incivility in the online public sphere. The aforecited research on political incivility suggests that controversial issues and events may lead ordinary citizens to express their opinions, lead to emotional engagement, and trigger uncivil expression (Theocharis et al., 2020). That is, fluctuations in incivility on social media may be triggered by offline events. Yet, because extant work mostly focuses on elections and/or specific short time periods, we do not know whether other events could lead to spikes in uncivil interactions online during non-election years and across different categories of online groups. Also, the implementation of various regulatory policies by social media platforms could be seen as an external event that influences users' behavior (Buntain et al., 2021). For instance, an analysis of YouTube's implementation of a policy regarding conspiracy-oriented channels showed a sharp and consistent change in trends of harmful content. Such policies serve to classify and regulate inappropriate behaviors and content, and may lead to an increase or decrease in incivility (Blackwell et al., 2017). We thus investigate the relation between online incivility and offline events, both socio-political and also platform specific. *RQ6: Have any specific external events triggered increases in incivility?*

2.2. Methods

2.2.1. Reddit

We rely on online behavioral data from Reddit, a social media platform with over 330 million users globally (Alexa, 2019) and 222 million in the US alone (Lin, 2021). Reddit is the only social media platform (apart from YouTube) that saw statistically significant growth since 2019 (Pew Research Center, 2021a) and steady growth in its user base since its inception. For example, from 2013 to 2019, the annual growth rate of monthly active users ranged from 21.42% (2014) to 47.06% (2017, Curry, 2021). Reddit is the ninth most visited website globally (Top, 2018), and the tenth most popular site in the US. Clearly, users' expressions therein are important to study.

As in other social media platforms, Reddit allows users to post content and discuss various issues in individual communities, which it calls "sub-reddits." A sub-reddit is a specific community dedicated to a particular topic where users can post a link, create a post, or comment on others' posts. Each sub-reddit has its own unique rules, moderators, and themes for submissions. Currently, there are more than 2.8 million sub-reddits, and more than 130,000 are active (receiving at least five comments a day, Lin, 2021). Those sub-reddits are of three privacy levels: public, restricted, and private. Any user can join and post in a public sub-reddit, but they can only join but not post in a restricted sub-reddit until the moderator approves. Private sub-reddits usually have rules governing admittance; users receive an invitation once they meet the admission requirements.

Several features of the platform are relevant to our focus on incivility. For one, unlike Facebook or Twitter, Reddit's core aspect is anonymity. Based on its privacy policy and its support for individual freedom of expression (Reddit, 2021), Reddit protects users' identities and does not require real-name or identity verification. Although this could result in uninhibited trolling, toxicity, or hate speech on the platform, Reddit has several mechanisms in place to prevent this from happening. Most sub-reddits have community guidelines developed by the creator and also moderators that explicitly forbid incivility, toxicity, trolling personal attacks, or other problematic languages in posts and comments. For instance, r/MachineLearning emphasizes “Be nice, no offensive behavior, insults, and attacks” as its first rule, and r/AskReddit also requires users to “be respectful to other users at all times and conduct your behavior in a civil manner.” The community rules are reinforced by both automatic tools called automods and human moderators. As a proactive tool, automods can remove and report posts and comments with inappropriate external links, words, and phrases. In addition, sub-reddit members are encouraged to report and downvote problematic posts and comments. Both auto and human reports go directly to sub-reddit moderators, who can remove the posts and comments that go against the sub-reddit’s rules and guidelines. In addition, administrators can remove content, ban users or even close down an entire sub-reddit based on their regular review of content and user reports.

Although Reddit had no specific anti-harassment policy prior to 2015, taking actions such as banning a user or taking down a sub-reddit only when certain concerns became public and received media attention (e.g., closing down of r/beatingwoman for violence against women and sharing users’ private information or r/TheFappening for posting hacked celebrity pictures), it announced its anti-harassment policy in May 2015. Reddit defined any behavior that makes users feel unsafe and shut users out of the conversation as uncivil (e.g., menacing someone and directing abuse at a user or a group). The then-developed user reporting system allowed human moderators and administrators to decide whether a comment and a user should be removed or prohibited (before that, users could only report content or groups by contacting administrators).

Furthermore, in 2019, Reddit invited bystanders (e.g., regular users not involved in the reported issues) to provide a third-person point of view on harassment reports. In addition, Reddit introduced machine learning tools to help organize and identify more severe cases. In 2020, in response to the George Floyd Protests, the policy was strengthened and further enforced. So far, Reddit still mostly relies on human judgment to identify any communities, users, or comments that go against its anti-harassment policy. Reddit's hands-off administration on the one hand and its gradually strengthened anti-harassment policy on the other hand make it a perfect platform to observe the natural flow of uncivil interactions.

We accessed all Reddit content from the beginning of Reddit.com (December 2005) up to December 2019 on PushShift's Reddit data using Google BigQuery[2]. In total, this yielded over 6.68 billion comments. Annually, the number of unique users commenting ranged from 23,793 (in 2006) to 80,788,041 (in 2019) (*M* = 19,401,466, *SD* = 26,733,025), and the number of comments ranged from 417,184 (in 2006) to 1,663,587,081 (in 2019, *M* = 477,154,362, *SD* = 526,611,725). In order to offer comprehensive evidence on the over-time fluctuations of incivility on Reddit, we identified the most popular sub-reddits, which represented 95% of the total Reddit comments each year. We did that by (1) the number of comments in the sub-reddit and (2) the number of users who posted in the sub-reddit. This has resulted in 9,355 sub-reddits that were most popular across the years. We therefore account for 95% of the entire Reddit universe. Among all identified sub-reddits, yearly comments in a sub-reddit ranged from 1,215

[2]. https://pushshift.io/using-bigquery-with-reddit-data/

(in 2006) to 84,457,656 (in 2019, *M* = 202,786, *SD* = 1,173,944), and yearly unique users ranged from 78 (in 2006) to 12,424,518 (in 2019, *M* = 33,162.3, *SD* = 161,665.3).

2.2.2. Sub-reddit Annotation

Our core questions pertain to the differences in incivility between political, non-political, and mixed sub-reddits, and also ideologically homogeneous (liberal or conservative) and heterogeneous (liberal and conservative) political and mixed sub-reddits. We eliminated non-English and banned sub-reddits and also those English-speaking sub-reddits that were specifically non-US (e.g., sub-reddits from or discussing Australia, Canada, India, or the UK),[3] resulting in 8458 sub-reddits for analysis (90.41% of all identified sub-reddits). We developed a coding manual to categorize each sub-reddit accordingly, as detailed below. Sub-reddits that discussed politics and news explicitly (e.g., r/politics, r/news) were categorized as political, while those revolving around non-political issues (e.g., r/nba, r/gaming) were categorized as non-political. In addition, the mixed category included sub-reddits whose purpose is not to discuss politics but where people discuss political issues (e.g., r/AskReddit, r/pics). In addition to these three categories, we classified the political and mixed sub-reddits into politically homogeneous or heterogeneous sub-reddits, and the former into liberal or conservative sub-reddits. Politically homogeneous sub-reddits were those where the majority of posts and comments were in favor of liberal/left/Democratic or conservative/right/Republican ideas, figures, and policies (e.g., r/BlueMidterm2018; r/Conservative, r/proguns). In turn, heterogeneous sub-reddits were those where posts and comments had mixed perspectives (e.g., some comments supporting and other

[3] As an additional exploratory analysis, we describe the aggregate over-time trends in incivility for the non-US, English speaking sub-reddits in Appendix E. We find the over-time trend and the proportion of incivility in all main categories was similar to those in the US sub-reddits.

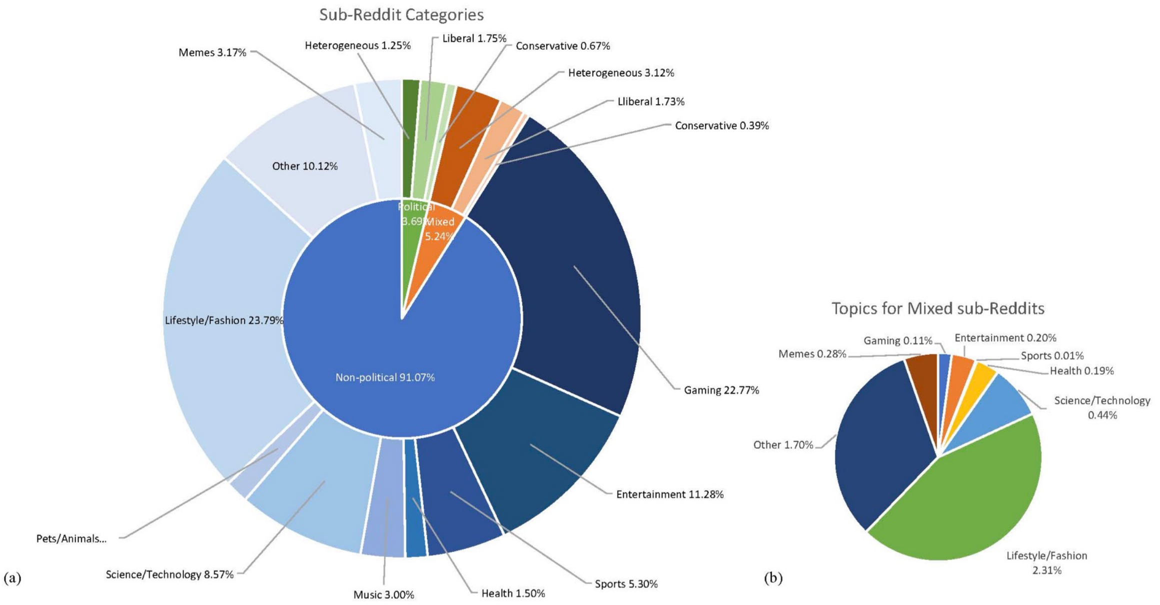
comments opposing the Democratic/Republican Party, or posts expressing both sides of an issue, such as r/news or r/PurplePillDebate). The politically homogeneous sub-reddits were further categorized as liberal (i.e., those supporting Democratic/liberal ideology and/or discussing socio-political issues from the Democratic/liberal perspective) or conservative (i.e., those supporting Republican/conservative ideology and/or discussing socio-political issues from the Republican/conservative perspective).

To address our question regarding incivility in *non-political* spaces, we also identified ten types of non-political sub-reddits based both on their overt purpose and content. Sub-reddits about games (video games, board games, etc.) and gaming services were categorized as *Games*; *Entertainment* category contained all sub-reddits about movies, TV programs, celebrities, and other entertainment; sub-reddits about sports, teams, and athletes were categorized as *Sports*; *Health* sub-reddits included all that discussed physical and mental health; *Music* category included sub-reddits discussing music, instruments, and musicians; *Technology* sub-reddits were those discussing science and technology developments and education; sub-reddits about pets and animals were categorized as *Pets/Animals*; *Lifestyle/Fashion* category contained sub-reddits about beauty, food, clothing, design, models, and lifestyle; and all sub-reddits dedicated to creating and sharing memes were categorized as *Memes*. The remaining sub-reddits were categorized as *Others*.

Seven trained coders labeled 8,458 sub-reddits (see Appendix A for a detailed coding procedure and inter-coder reliability). Figure 2.1 shows the distribution of the categories and Appendix B presents specific examples. We identified 312 political sub-reddits (3.69% of total sub-reddits), of which 66.03% (206) were politically homogeneous and 33.97% (106) heterogeneous. Among the homogeneous sub-reddits, 72.33% (149) were liberal and 27.67%

(57) conservative. Further, 443 sub-reddits were classified as mixed (non-political with at least 40% of posts and/or comments pertaining to politics; these comprised 5.24% of total sub-reddits). Among the mixed sub-reddits, 40.41% (179) were ideologically homogeneous (146 liberal, 33 conservative) and 59.59% (264) heterogeneous. The remaining 7703 sub-reddits were non-political (91.07% of total sub-reddits), with the largest groups of non-political sub-reddits being lifestyle and fashion (n = 2012, percentage = 23.79%), followed by games (n = 1926, percentage = 22.77%) and entertainment (n = 954, percentage = 11.28%).

Figure 2.1. *Sub-reddit Categories with Percentage.*



Note. (a) The inner circle represents the distribution of main categories: political, mixed, and non-political with percentage of total number of sub-reddits. The outer circle represents the second level categories for each main category with percentage of total sub-reddits. The second level categories include heterogeneous, liberal, and conservative for political sub-reddits; heterogeneous, liberal, and conservative for mixed sub-reddits; gaming, entertainment, sports, health, science/technology, lifestyle/fashion, memes, and others. (b) The circle represents the topics in mixed sub-reddits with the percentages of total number of sub-reddits, including gaming, entertainment, sports, health, science/technology, lifestyle/fashion, memes, and others. In addition, out of political sub-reddits, heterogeneous takes 33.97%, liberal is 47.76%, and conservative is 18.27%. For mixed sub-reddits, 59.59% are heterogeneous, 32.96% are liberal, and 7.45% are conservative.

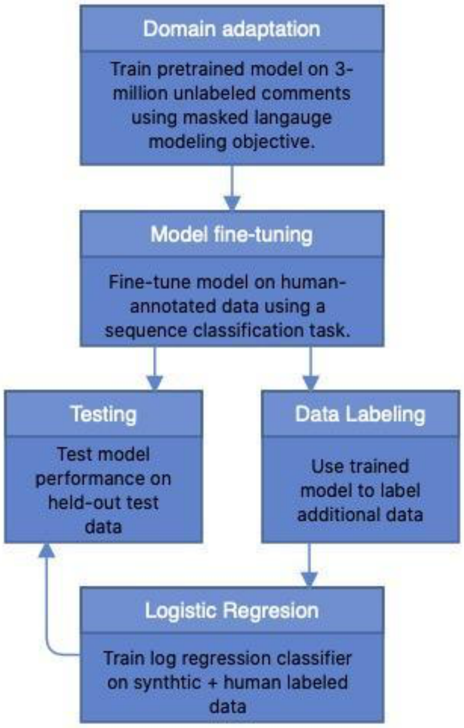
2.2.3. Incivility Annotation and Classifier

To classify Reddit content as uncivil or not, we developed and validated an incivility classifier. A coding manual was developed based on previous research (Coe et al., 2014), and three new trained coders labeled Reddit comments with binary labels as civil or uncivil. Uncivil comments were those that included (1) name-calling, mean-spirited or disparaging words directed at a person, or a group of people; (2) aspersion, mean-spirited or disparaging words directed at an idea, plan, policy, or behavior; (3) vulgarity, profanity or language that would not be considered proper; (4) pejorative or disparaging remark about the way in which a person communicates. For instance, comments such as “It's OK, you'll hit puberty one day” or “you’re a dumbass for simplifying the issue and trying to jump right into the helm of the ‘y’r all hypocrites’ bandwagon” were coded as uncivil. Our approach accounted for both the content as well as the targets of incivility to create a comprehensive dataset for model building. Coders received five runs of coding exercises, with overall inter-coder reliability resulting in a Fleiss’s kappa of 0.663, and then moved on to individual coding. A final set of 4,000 stratified sampled comments from each year was randomly assigned to coders, and the individual coding and training coding were together used for supervised model building.

In order to automatically identify incivility, we decomposed the task into three steps. We first developed neural binary classifiers built on top of large transformer-based language models, namely BERT (Devlin et al., 2018). First, the pre-trained BERT model was further pre-trained for domain adaptation on 3 million unlabeled Reddit comments using a masked language modeling objective. Then the model was fine-tuned for four epochs on 5000 human-labeled comments with 10% of the data set aside for training validation and 1000 coded comments set aside for model testing (see Davidson et al., 2020). The final *F1*-score[4] for the classification model was 0.786. Next, we tried to improve computational performance by utilizing DistilBERT (Sanh et al., 2019), a more compact version of BERT trained using a model distillation technique. The final F1-score of the DistilBERT model was 0.802. Considering the large scale of our dataset, using BERT or DistilBERT models to classify more than ten years' Reddit data would be both time-consuming and computationally expensive. To address this constraint, in the third step, a logistic regression classification model was trained using 5 million Reddit comments labeled by our fine-tuned DistilBERT model, in addition to the smaller human-annotated dataset. The final logistic regression model achieves an *F1*-score at 0.779 which is similar to the performance of our BERT and DistilBERT models, and our model error falls within the 95% CI of [0.0297, 0.0547]. Figure 2.2. gives an overview of the computational framework. and Appendix C offers detailed model building procedures.

Figure 2.2. *Computational Framework Flow Diagram.*

[4] F1-score is a measurement of model accuracy for binary classification, which is calculated from precision and recall. Precision is the number of true positives (the incidents which are 1 and also identified as 1) divided by the number of all positives, while recall is the number of true positives divided by the sum of true positives and false negatives (the incidents which are 1 but identified as 0 by the machine). F1-score is the harmonic mean of precision and recall.


Domain adaptation
Train pretrained model on 3-million unlabeled comments using masked langauge modeling objective.
Model fine-tuning
Fine-tune model on human-annotated data using a sequence classification task.
Testing
Test model performance on held-out test data
Data Labeling
Use trained model to label additional data
Logistic Regresion
Train log regression classifier on synthtic + human labeled data

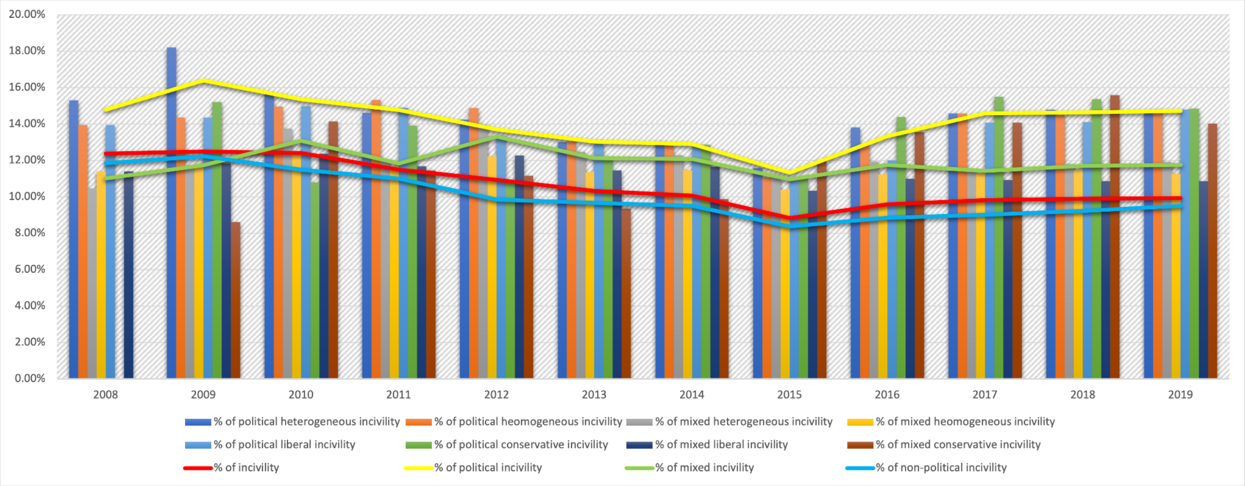
Note. The graph depicts the steps of the machine learning process. Both BERT and DistilBERT followed the first two steps, but only DistilBERT was used for data labeling. The generated labeled data from the DistilBERT model was used for logistic regression.

2.3. Results

2.3.1. Macro Trends

The overall yearly trends in the prevalence of incivility relative to the total content contributed are shown as the red line in Figure 2.3. Between 2008 and 2019, total incivility - depicted with the red line - fluctuated between 8% and 12%, an estimate that is largely consistent with evidence from Twitter (Siegel et al., 2018; Theocharis et al., 2020). After slight decreases in the general proportion of incivility until about 2015, when the total proportion of comments classified as uncivil reached the lowest point of 8.84%, incivility has been gradually increasing since, with its levels rising to around 10% in 2016 and 2017. We note, however, that this increase was not dramatic and that the proportion of comments categorized as uncivil did *not* return to the high, pre-2015 levels around 12%, which is when Reddit initiated its anti-harassment policy and banned several sub-reddits promoting incivility and hateful speech. Addressing RQ2.1, we note that the proportion of incivility fluctuates only slightly, with a current upward trend, and can be affected by the policies of social media platforms.

Figure 2.3. *Yearly Incivility Proportion from 2008 to 2019.*



Note. The percentage shown in the graph is the proportion of incivility content over relative total content.

2.3.2. Political, Mixed, and Non-political Discussions

Are there variations in incivility across political, mixed, and non-political groups? Addressing RQ2.2, *political groups* - the yellow line in Figure 2.3. - contain the highest proportion of incivility among the three major categories (i.e., political, non-political, mixed) across all the years analyzed, with the percentage oscillating between 10% and 17%. Results from one-way ANOVA (F = 32.095, $p < 0.001$) showed a significant difference among these categories, and post hoc Tukey's HSD indicates that incivility in political groups is significantly higher than in mixed (diff = 0.025, $p < 0.001$) and non-political groups (diff = 0.040, $p < 0.001$). It is in the political groups that we observe the steepest increase in incivility after 2015, likely due to the highly contentious 2016 presidential elections.[5] Incivility in political groups increased by 33.12% between 2015 and 2017 (see also Nithyanand et al., 2017b) and has been growing gradually since 2017.

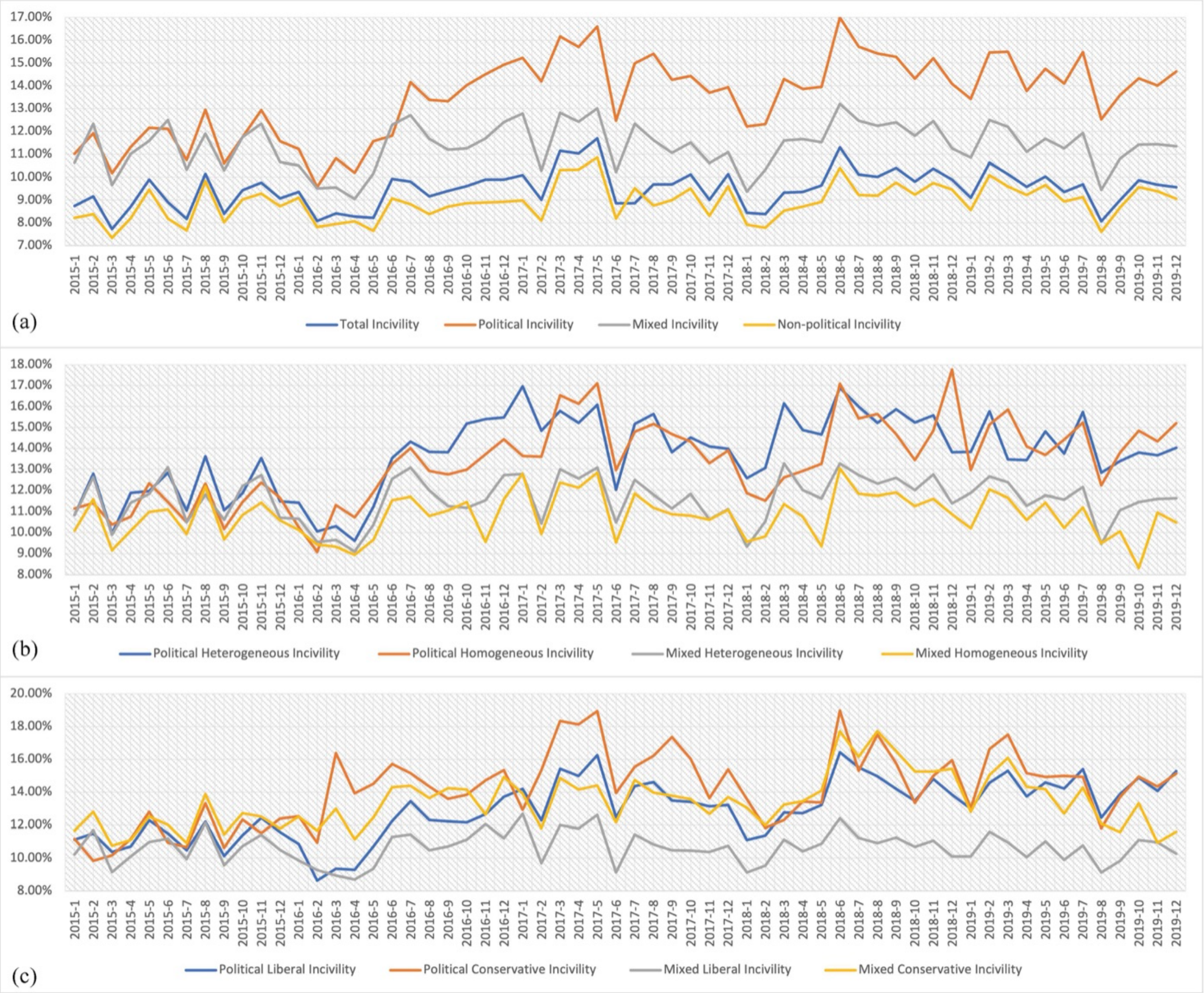
In *mixed* groups, where participants discuss political and non-political issues, the proportion of incivility ranged between 11 and 13%, which is significantly higher than in non-political groups (t = -2.940, df = 12, p < 0.01) and significantly lower than in political groups (t = 5.333, df = 12, p < 0.001). But at some points, such as during the 2015-2016 period, incivility in mixed groups spiked, reaching levels of incivility similar to that in political groups. The temporal variations in incivility in *non-political* groups are similar to the total trends in incivility and those in political sub-reddits. As could be expected, the proportion of incivility to overall content in these groups is significantly lower than overall proportion across all the sub-reddits (t = 8.045, df = 12, p < 0.001), as well as that in political and mixed groups.

To shed more detailed light on the variations between 2015 and 2019, we also analyzed monthly data. Figure 2.3(a) depicts large variations in incivility, yet the relative proportion of incivility in general and also in mixed and non-political groups remains stable. Consistent with the yearly trend, the incivility proportion in political groups (F = 2048.534, p < 0.001) is significantly higher than in mixed (diff = 0.065, p < 0.001) and non-political groups (diff = 0.064, p < 0.001). Notably, however, political incivility increased with several spikes. The peak in July 2016 can be linked to the 2016 Democratic National Convention and the early email leak of the Democratic National Committee (which can be confirmed by the boost of incivility in the liberal group, See figure 2.3(c)). The observed spike in May 2017 overlapped with several actions related to the investigation of Russian interference in the 2016 US election, including the

[5] Incivility also peaked in 2009. Based on the examination of a random sample of uncivil political comments in 2009, this increase may be due to discussions of equality and medical care.

Great America Committee and the dismissal of James Comey. The last peak in June 2018 corresponds to several protests against the family separation policy.

Figure 2.4. *Monthly Incivility Proportion from January 2015 to December 2019.*



Note. The percentage shown in the graph is the proportion of incivility content over relative total content. (a) is the proportion of incivility overall and in three main categories; (b) is the proportion of incivility in political heterogeneous/homogeneous and mixed heterogeneous/homogeneous groups; (c) is the proportion of incivility in political liberal/conservative and mixed liberal/conservative groups.

2.3.3. Ideological Homogeneous vs Heterogeneous Political and Mixed Discussions

RQ2.3 asked about incivility in ideologically homogeneous vs. heterogeneous groups. Figure 2.3, which also summarizes the trends for political and mixed categories of ideologically homogeneous/heterogeneous groups from 2008 to 2019 in bars, shows that incivility in all four categories oscillated between 10 to 20%, reaching its lowest levels in 2015, and gradually increasing since then. Although incivility in ideologically *heterogeneous political* groups, where users encounter others with differing opinions, was higher than incivility in ideologically *homogeneous* political groups, Welch's t-test showed this difference is not significant ($t = 1.545$, $df = 11$, $p = 0.15$). A detailed monthly trend from 2015 to 2019, shown in Figure 2.4(b), shows small yet growing fluctuations in the proportion of incivility in ideologically homogeneous and heterogeneous political groups. Two noticeable spikes in ideologically *heterogeneous* groups, which did not have corresponding spikes in the *homogeneous* groups, occurred in January 2017 and March 2018. The former spike overlapped with executive order 13769 (also known as Muslim Ban) and Trump's inauguration. The spike in March 2018 could be attributable to the breaking news of Cambridge Analytica's involvement in Trump's presidential campaign. In turn, there were two spikes in incivility in ideologically *homogeneous* political groups that did not occur in ideological *heterogeneous* groups, in December 2018 and March 2019. The former can be linked to the longest US government shutdown in history and the latter was due to the release of the Mueller Report about Russian interference in the 2016 election.

Incivility in *mixed* groups that were ideologically *heterogeneous* spiked in 2010 and 2012 and reached its lowest levels in 2015. Sampling comments from mixed sub-reddits suggest that the reasons for these spikes were discussions about healthcare reform in 2010, whereas the peak in 2012 was due to the presidential election. In turn, incivility in *mixed* groups that were ideologically *homogeneous* fluctuated within a small range of 10 - 12%, significantly lower than that in *mixed heterogeneous* groups ($t = 2.881$, $df = 11$, $p < 0.05$). Furthermore, the monthly proportion of incivility in ideologically *heterogeneous* mixed groups from 2015 to 2019 was also significantly higher than that in *homogeneous* mixed groups ($t = 9.439$, $df = 59$, $p<0.001$), indicating - again - greater usage of incivility in ostensibly non-political groups where both liberals and conservatives sometimes discuss politics.

To answer RQ2.3, ideologically *heterogeneous* mixed groups entail more incivility than *homogeneous* groups, but it is not the case that discussions in ideologically *heterogeneous* political groups are necessarily more uncivil than discussions in ideologically *homogeneous* political groups. In addition, discussions in ideologically heterogeneous *mixed* groups were significantly less uncivil than those in ideologically heterogeneous *political* groups ($t = -12.987$, $df = 59$, $p < 0.001$), confirming recent findings (Rajadesingan et al., 2021).

2.3.4. Conservative vs Liberal Homogeneous Groups

Next, we examined the fluctuations in incivility in ideologically homogeneous, liberal, or conservative political and mixed sub-reddits. The bars in Figure 2 show that incivility was rather stable in liberal groups - especially mixed - as compared to conservative groups. Between 2008 to 2015, incivility in homogeneous *liberal political* groups gradually decreased from 14.98% to 11%, and then returned back to 14.97% in 2019. In turn, the proportion of incivility in homogeneous *liberal* groups that were *mixed* (discussing non-political issues but sometimes diverting to politics) decreased before 2015 and remained stable at around 11%, suggesting that the effects of the anti-harassment policy initiated by Reddit were especially effective in liberal mixed groups (perhaps because these groups were victims of disproportionate amount of trolling and harassment prior to the policy).

When it comes to incivility in ideologically homogeneous *conservative* groups, the oldest identified conservative sub-reddit was founded in 2009. Before 2015, the proportion of incivility in *conservative political* groups reached two peaks in 2011 and 2014. The peak in 2011 may be linked to the nomination for the 2012 presidential election. In turn, uncivil comments from 2014 were mostly about gun control, minority groups, income equality, and climate issues, probably reacting to offline events such as the legalization of same-sex marriage in several states, the raising of the minimum wage, and news about mass shootings and gun laws. A rapid increase in incivility in *conservative political* sub-reddits occurred after 2015 and gradually declined after 2017, consistent with the findings of political incivility during the 2016 election period (Nithyanand et al., 2017b). In *conservative mixed* groups, incivility peaked in 2010 and declined to 9.35% in 2013 and then slowly climbed back to 15.58% in 2018. *Conservative mixed* groups were least affected by Reddit's anti-harassment policy, as incivility around 2015 was not at its lowest point.

To answer RQ2.4, we see that before 2015, incivility in *liberal* political and mixed groups was higher than in *conservative* political and mixed groups. The trend reversed after 2015, with the highest incivility proportion in *conservative political* groups, followed by *conservative mixed* groups, *liberal political* groups, and *liberal mixed* groups. In fact, Welch's t-tests using monthly trends from 2015 to 2019, shown in figure 2.4(c), confirm that the proportion of incivility in *conservative political* groups was significantly higher than in liberal political groups (t = -6.304, df = 59, p < 0.001). Incivility in *conservative mixed* groups was also significantly higher than in liberal mixed groups (t = -16.049, df = 59, p <0.001). Given that conservative news media are more likely to use outrage and divisive language (Sobieraj & Berry, 2011), this difference could be a reflection of the mainstream political discourse.[6]

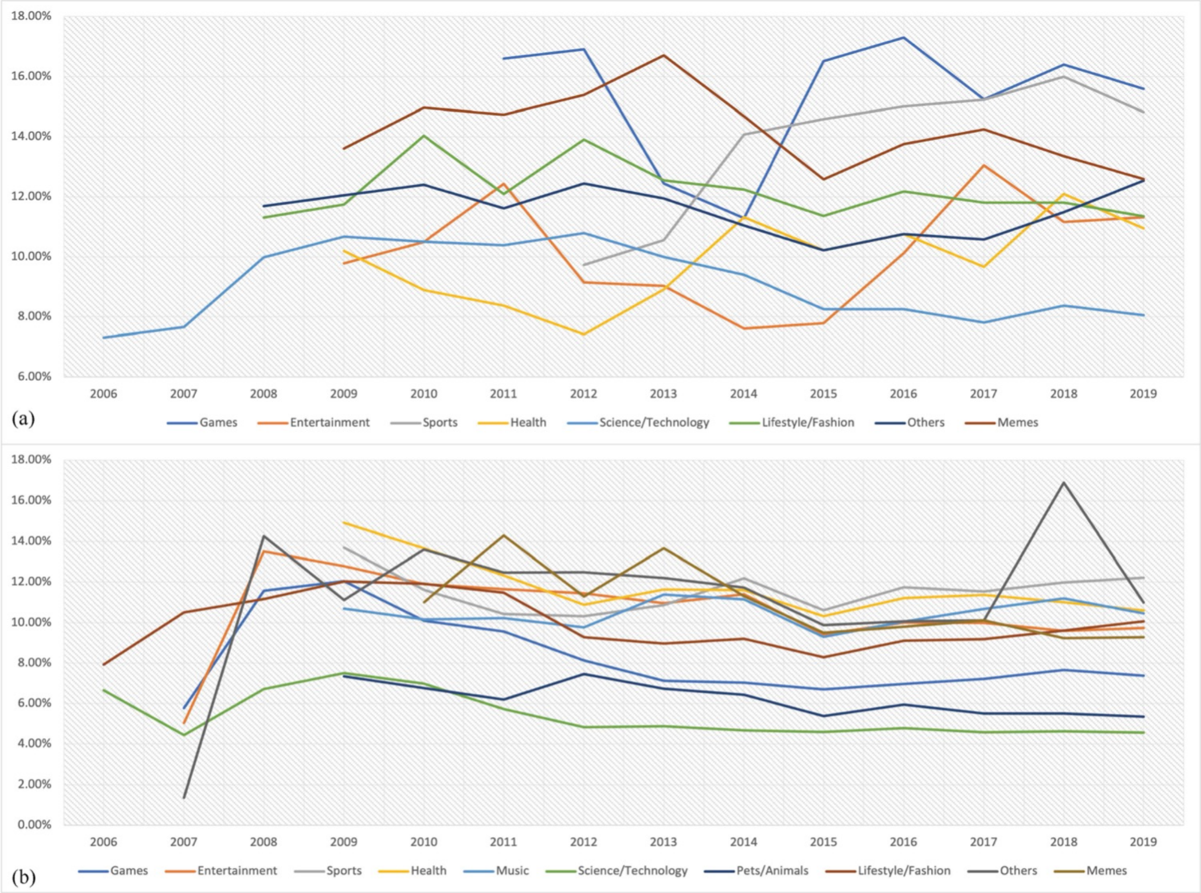
[6] When using the *yearly* data from 2008 to 2019, we find no significant differences in incivility between liberal and conservative groups, both political (t = 0.403, df = 10, p = 0.695) and mixed (t = -1.005, df = 10, p = 0.339).

2.3.5. Topics in Mixed and Non-political Groups

To answer RQ2.5, we turn to the non-political topic categories in mixed and non-political sub-reddits. Incivility in *mixed* discussions varied across topics and years (shown in Figure 2.5(a)).[7] Discussions about games, sports, and memes were most uncivil, perhaps because games and sports are ego-involving and, similarly to politics, generate the us-versus-them divide. In contrast, health and science/technology sub-reddits were the least uncivil, the former likely because many health sub-reddits discuss marijuana legitimization, which is supported by most participants, and the latter likely because most science and technology sub-reddits focus on problem-solving, which again, does not generate high incivility. When it comes to incivility in *non-political* sub-reddits, shown in Figure 2.5(b), discussion about science and technology and pets and animals were most *civil*, whereas sports, memes, and entertainment generated more incivility on average.

Figure 2.5. *Yearly Incivility Proportion in Mixed and Non-political Topics from 2006 to 2019.*

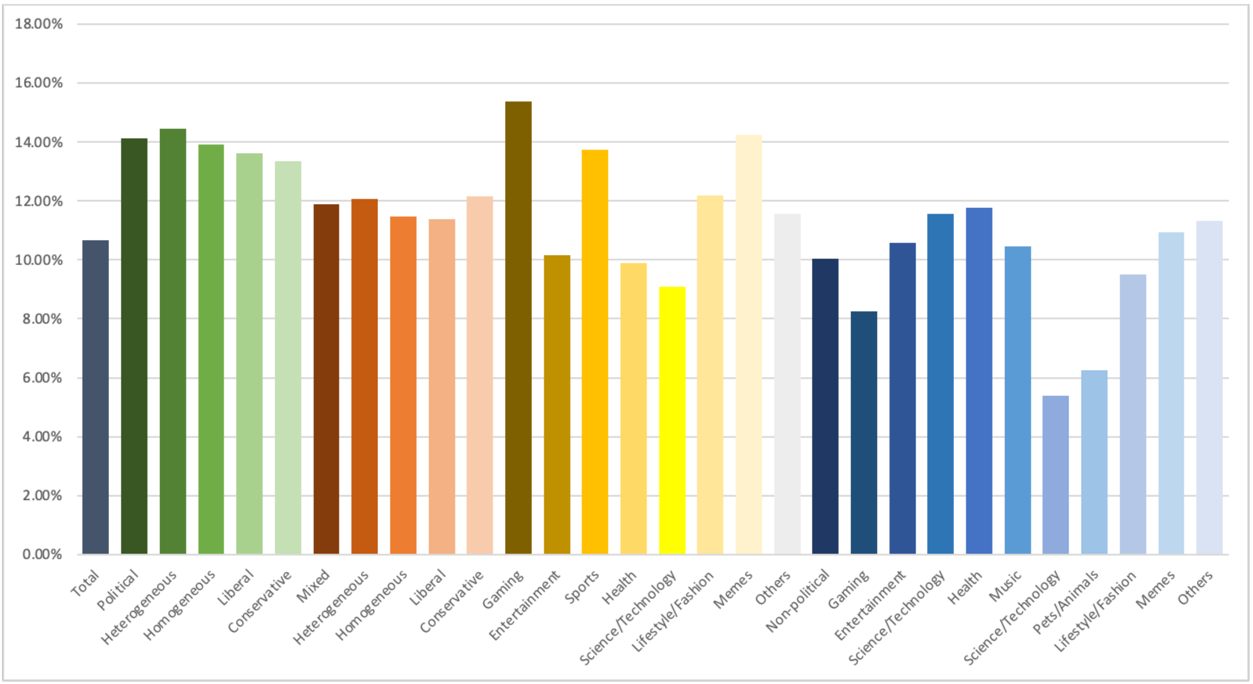
[7] Although ten distinct categories were identified for *non-political* sub-reddits, only eight were found in *mixed* groups (and are used for the comparison; i.e., the category of Music and Pets/Animals was not present in mixed groups).



Note. The percentage shown in the graph is the proportion of incivility content over relative total content. (a) is incivility in mixed topics; and (b) is incivility in non-political topics.

Lastly, we calculated the average incivility proportion for all categories from 2008 to 2019, shown in Figure 2.6. (See Appendix F for detailed statistics). Interestingly, the highest average proportion of incivility was found in the mixed gaming category, higher even than in political sub-reddits (Welch's t-test confirmed the significance of the difference, $t = 4.758$, $df = 59$, $p < 0.001$). Online gaming communities have their unique culture that often validates disparaging or disrespectful language, leading to this high aggregate proportion. Additionally reviewing comments from the mixed gaming sub-reddits suggests that those were mostly massive multiplayer online games with international servers that require a high level of communication among players and that are competitive by design, providing a hotbed for uncivil discourses.

Figure 2.6. *Average yearly Incivility Proportion for All Categories from 2008 to 2019.*



Note. The percentage shown in the graph is the proportion of incivility content over relative total content.

2.3.6. External Events

The fluctuations of incivility in different categories were addressed throughout above, with some spikes in response to offline events and the changes in platform policies. Incivility in political and mixed groups tends to surge around highly contentious political events, such as election campaigns (e.g., 2016 presidential election), political scandals (e.g., Cambridge Analytica), and controversial orders (e.g., Muslim ban). Incivility in non-political groups also shows some spike during offline events such as sports events (e.g., 2019 super bowl), industrial scandals (e.g., 2016-2017 US gymnastic sexual abuse scandal), and industrial controversies (e.g., 2016 complaints about sexualized characters in games). Furthermore, there was a sharp decrease in incivility in all categories except mixed conservative groups in 2015, which corresponds to Reddit's anti-harassment policy. Consistent with the findings of previous research (Buntain et al., 2021), incivility in most categories after 2015 remained at a stable level, suggesting the intervention has both immediate and long-lasting effects.

2.4. Discussion

Even though incivility is a growing concern for the public, politicians, and social media platforms, we know relatively little about its fluctuations over time and its prevalence across different types of online discussions. This study offers this key descriptive evidence, showing how incivility developed over time on Reddit in political, mixed, and non-political groups, and also whether and how it differed in each group. We relied on a combination of machine learning methods and traditional statistical inference to examine the dynamics of online incivility on Reddit, an increasingly popular social media platform (Pew Research Center, 2021a).

Our findings suggest that extant worries about the prevalence and rapid growth of online incivility may have been overstated. Incivility is not ubiquitous in Reddit discussions and has not dramatically grown in recent years with the help of platform interventions. Its proportion is rather consistent, oscillating between 8% and 12 %. The illusion of ever more incivility may be due to the increasing volume of total online discussion in general, yet - again - the proportion of incivility to this overall volume of content is relatively stable. We also note that even though Reddit could invite greater incivility than Facebook or Twitter, due to its largely anonymous nature, the estimates in our data are largely similar to those from studies of other social media platforms (Siegel et al., 2018; Theocharis et al., 2020).

Our other noteworthy findings relate to the differences in incivility across different categories of online discussion spaces. For one, consistent with anecdotal observations, users encounter more name-calling and disparaging or vulgar language in online discussions revolving around politics. That is, incivility is higher in political groups, followed by mixed groups whose focus is not politics, but which nevertheless entail socio-political discussions, and then non-political groups, where users discuss politics only rarely, if at all. A notable exception to this overall pattern is the mixed gaming category, where the aggregate proportion of incivility across all the years is *higher* than in political groups. Unlike other mixed groups, where incivility may be closely moderated and restricted by group members, gaming groups are known for endorsing incivility as a special social norm and encouraging uncivil behaviors such as flaming and trolling (Shen et al., 2020). Thus, incivility is likely to be promoted in such groups no matter whether discussions revolve around games or politics.

Second, even though ideologically diverse political discussions are seen as the breeding ground for uncivil discourse (Rossini, 2020), political sub-reddits involving participants expressing liberal and conservative perspectives are not necessarily more uncivil than ideologically homogeneous political groups. Furthermore, it is the ideologically heterogeneous mixed groups, where discussion about political issues may be unexpected and/or auxiliary and which involve diverse discussants, that entail *less* incivility than ideologically heterogeneous political groups, the sole purpose of which is to discuss politics. That is, heterogeneous political discourse is *less* uncivil in mixed sub-reddits than in political sub-reddits, consistent with the findings of uncivil cross-partisan discussions in non-political versus political online spaces (Rajadesingan et al., 2021). It is possible that political discourse is carefully moderated and restricted by moderators and members in sub-reddits that are designed for non-political topics, thereby preventing incivility. Alternatively, it may be the case that once users establish common ground on non-political topics (e.g., as chihuahua owners or Kardashians' fans), political disagreements with dissimilar discussants does not generate the same levels of emotional response, and thus incivility, as political disagreement in groups solely dedicated to current events and potentially divisive policies. Even though our large-scale project cannot speak to the underlying mechanisms, our findings clearly suggest that the dynamics of political discourse online are contingent on social context, such that differences in the types of conversation lead to different expressions of incivility.

Third, platform-specific, as well as exogenous factors, may powerfully shape online discourse, trends in incivility included. With regard to the former, the presented patterns underscore the effectiveness of anti-harassment policies by social media platforms. In 2015, when Reddit allowed its users to report abuse and harassment and consequently banned sub-reddits promoting racism or antisemitism, overall incivility on the platform dramatically declined. In turn, underscoring the influence of the overall divisive political environment on online discussions in the subsequent years, we show that incivility clearly increased around the contentious 2016 elections and during Trump's presidency. Also, external socio-political events such as debates about welfare, gun control, or sexual minorities, also led to fluctuations in incivility, in line with previous research about political events impacting the temporal dynamics of incivility (Theocharis et al., 2020). When these external events are divisive or controversial, peoples' expressions and exchanges on social media may get heated and uncivil.

In fact, we note that after 2015, incivility in *political* groups increased at higher rates than on the platform in the aggregate and that both political and mixed *conservative* groups generated significantly more incivility than liberal groups. This suggests the differential effects of the political environment. It could be that conservatives were more susceptible to the polarized context, especially during the presidency of Donald Trump, known for his device and often inappropriate rhetoric, which could have ‘trickled down’ to online communities on Reddit. In a related vein, this difference could be a result of conservatives consuming news from conservative sources known for their inflammatory expression (Sobieraj & Berry, 2011). Picking up the elite cues, either from politicians or news media, conservative Reddit users could be adopting certain expressions in their political discourse or using it as a basis for online discussion.

When interpreting these findings, a few limitations of our project should be kept in mind. First, though the data was collected as 95% of total universe of Reddit, the data will never be perfect. As we cannot access the deleted or removed comments, our results may be biased. Especially with the implementation of anti-harassment policy after 2015, the extremely uncivil comments were highly likely to be removed or deleted. The observed cases of incivility may be milder, which may have led to underestimations of incivility on Reddit both on the level of severity and amount. The observed stability of incivility might be the result of the platform interventions, which again, emphasizes the crucial role of platform policies in providing a welcome and deliberative environment for discussion. Without platform interventions, we are likely to observe an increased proportion of incivility. Additional analysis with the total number of deleted and removed comments is shown in Appendix E.

Second, future work should apply a more sophisticated classification of political, non-political, and mixed groups, using machine learning applied at post level to automatically detect whether the discourse is political only, non-political only, or both. Also, our findings may not generalize to other social media platforms, such as Facebook or YouTube. Reddit has a unique culture and is known for its grassroots - as opposed to algorithmic - moderation system. As such, the uncivil discourse patterns observed on Reddit may not be found on other platforms. The fact that our estimates are largely similar to those detected on Twitter (Theocharis et al., 2020), suggests certain robustness to our findings. Yet naturally, a systematic cross-platform work would be an important addition to the literature. Perhaps most importantly, our analysis only takes into account users' posts or comments, i.e., textual expressions. As such, we lose the information conveyed via memes, pictures, and videos.

Despite these limitations, our research is the first to offer systematic descriptive evidence of temporal dynamics of incivility on Reddit, over 11 years, across various categories of discussions, and focusing on thousands of sub-reddits that account for 95% of users and comments over this time period. Perhaps counter-intuitively, the rise in incivility has not been as steep as many observers fear and continues to constitute a similar fraction of the overall online discussion (so naturally increasing in total, but not proportionally), with some important variations across different contexts of the overall online public sphere. We hope that future work addresses these different dynamics and mechanisms, shedding more detailed light on the role of group culture, topical influence, offline socio-political events, platform-level interventions, such as reporting or moderating systems, and the users themselves. All these macro-, meso-, and micro-level factors influence incivility and need to be accounted for conceptually and analytically. Inasmuch as name-calling, disparaging and vulgar language, and other personal attacks have negative effects on public discourse, online discussions, social media users, and social media platforms themselves (Liang & Zhang, 2021), these investigations are important. As social media platforms have become a major source of news and information and an important channel for political discussion, understanding the complexity of online incivility is the necessary first step to promote a healthy dynamic of political deliberation in contemporary democracies.

Chapter 3. Who Would Respond to A Troll? A Social Network Analysis of Reactions to Trolls in Online Communities

Online anti-social behaviors such as trolling, cyber-bullying, and flaming (Hardaker, 2010; Lea et al., 1992; Smith et al., 2006) have increased alongside the volume of internet users and online interactions. According to a 2018 national survey conducted by Statista, 53% of US adults have been subjects of any online harassment (Clement, 2019), and more than 60% have witnessed trolling on social media or video/vlog platforms daily (Kunst, 2019). Trolling behavior is considered anti-social behavior that provokes emotional responses and disrupts on-topic discussions in online communities (Shin, 2008). It not only results in the dysfunction of online communities (Ybarra et al., 2006), but it can also lead to serious emotional distress and offline violence (Hinduja & Patchin, 2007).

Previous work mostly focused on the identification and motivation of trolling behaviors (Coles & West, 2016; Engelin & De Silva, 2016; Hardaker, 2010; Lökk & Hallman, 2016). A wide variety of motivations for trolling behaviors exist. They can be as complex as manipulating public opinions (Engelin & De Silva, 2016) or as simple as a desire to attract attention (Herring et al., 2002). However, not much work has focused on victims of trolling behaviors. Unlike other anti-social behaviors, such as cyberbullying, which have specific targets or victims, trolling behaviors typically do not purposefully target specific victims. Rather, trolls indiscriminately harass everyone and anyone provoked by their behavior (Hardaker, 2010). Once provoked, the victims may experience negative emotions and antagonism (McCosker, 2014). The detrimental impact of trolling on communities are varied across audiences and contexts, including failure to discuss meaningful topics in a feminist forum (Herring et al., 2002), increased self-censoring within minority groups on Twitter (Olson & LaPoe, 2017), depression, hostility, and interpersonal sensitivity among college students who are trolling victims (Crosslin & Golman, 2014), and increased delinquent behaviors offline among adolescent victims (Hinduja & Patchin, 2007). Some people are easily triggered by trolls, while others are more resistant. With all community members as trolls' potential victims, trolling behaviors also have consequences for the entire community, with some communities being more solid than others. How community members react to trolls and how a community as a whole reacts to trolls are of vital importance to cope with the consequences of trolling behavior. But research remains limited regarding who and which communities are more vulnerable to trolls.

This research fills an important gap in online trolling research on how online communities' members process and react to trolls. Drawing from social identity theory, we conceptualized community engagement and exchanges among community members as acts to maintain social identity in the face of threat. We analyzed the change in social network dynamics with trolls structurally embedded in the network of a community and leveraged network analysis to provide insights on the dynamic of trolling in online communities.

By analyzing commenting data from 23 YouTube videos, we illustrated that the structural position of a community member in a community network provides useful information on how they would react to trolls as a way to defend their social identity. Network structural position provides information not only about the community dynamic but also about individual members' involvement, emotional devotion, and attachment, which can form and reinforce the social identity that predicts how they will react to trolls (Cover, 2012). We found that people in a more central position would be more likely to reply to trolls, especially in dense communities and when there are prior interactions among community members and trolls.

3.1. Theory and Hypotheses

3.1.1. Definitions of Trolling

Previous literature has defined trolling behaviors in many different ways. One stream of research focuses on the deceptive aspect of trolling behavior. An early study from Donath (2002) on trolls in USENET newsgroups focused on deception and manipulation. Donath (2002) stated that trolls are troublemakers who pretend to be legitimate participants, provoking others by providing incorrect advice and damaging their trust. Engelin and De Silva (2016) defined trolling as "interrupting, harassing, or trying to impose opinions to others" (Engelin & De Silva, 2016, p. 4) by using fake personas on Twitter. In some cases, trolls are considered hackers, who use deceptive comments to find valuable information about the victims to attack them on Reddit (Breeze, 2012). Those definitions all indicate that deception is one of the key characteristics of trolling and that trolling messages may contain misinformation, logical fallacies, or misleading information. In those cases, the intention of trolls is to manipulate public opinion. However, deception is not the only defining feature of trolling behavior, and the intention of trolls may vary as well.

Another way to define trolling behavior is through its provocative nature. These definitions include "causing disruption and/or triggering or exacerbating conflict" (Hardaker, 2010, p. 237); intensifying engagement for the purposes of the trolls' own amusement (Nevin, 2015); transgressions of community norms that result in anger, harm or discomfort (Bergstrom, 2011); and "posts of erroneous or inflammatory information with the intention of provoking a strong reaction" (Merritt, 2012, p. 2). These definitions emphasize the results of trolling behaviors that include negatively affecting people's internet experience, reducing the chance of problem-solving, and misleading on-topic discussion.

Other scholars comprehensively define trolling by combining deception and provocation (Klempka & Stimson, 2014). Herring et al. (2002) defined trolling as “messages from a sender who appears outwardly sincere; messages designed to attract predictable responses of flame; and messages that waste a group’s time by provoking futile argument” (p. 375). They can be outright swearing, personal attacks, veiled insults, sarcasm, and off-topic statements (Cheng et al., 2017). (Hardaker, 2010) defined trolling behaviors as deception, aggression, disruption, and success. According to Hardaker (2010), a trolling message with negative emotion and false information will lead to dysfunction of the discussion or an unpleasant experience.

Existing research shows that trolling is a complex behavior with multiple intentions and different results. In this study, we adopt the comprehensive perspective, as it considers both the characteristics and the outcome of the behavior. We consider trolling behavior as an aggressive behavior with the intention to disrupt the discussion and displease others, and it usually manifests as providing fallacy regarding the topic being discussed, making a snap judgment of others or attacking people, and using profanity.

Note that after the 2016 US presidential election, the term “Russian trolls” took on a different meaning in the public discourse, one that is distinct from the definition reviewed above. In that context, trolls were defined as an ideological device (Fichman & Sanfilippo, 2016), paid astroturfers (Zelenkauskaite & Niezgoda, 2017), or bots (Alsmadi & O'Brien, 2020) that express an extreme political opinion to purposefully mislead people. This paper follows the original definition of trolls (Herring et al., 2002) based on their deviant behaviors, regardless of political motivation.

3.1.2. Understanding Reactions to Trolling through Social Identity Theory

Social identity is an individual's self-concept generated from perceived membership in a relevant social group with emotional and value significance to individuals in the social group (Hogg, 2016). It is dynamic, self-reflective, and performative (Greenhow & Robelia, 2009), and it is a social product of a given social environment and context (Zhao et al., 2008). Social identity theory (SIT) states that individuals' behavior within a group is largely determined by the perceived social identity, specifically, to achieve or maintain a desirable positive social identity. In an online community, the social identity motivates members to spend time in the community, interact with other community members, and share knowledge, which maintains a positive self-defining relationship with other members and reinforces the attachment (Shen et al., 2010), and also facilitates the emergence of collective behaviors (Ackland & O'neil, 2011). Previous research showed that an online community member's social identity is positively related to their knowledge contribution to the community (Shen et al., 2010). The shared knowledge and social resource produce shared social characteristics that lead to identification, as the understanding of conventions in the group for doing particular social acts is also shared (Ochs, 1993). SIT helpfully explains the mechanisms within individuals' online social engagement and their relationships with other internet users.

Online community members react to trolls in a variety of ways. As shown in previous research, these reactions generally include leaving the community altogether, actively engaging with the trolls, and ignoring the trolls. Ansong et al. (2013) looked at trolls in Ghanaian WhatsApp groups and noticed that trolling behaviors caused some legitimate community members to exit the communities. Research conducted on the political discussion in USENET shows that trolls are isolated and relegated to the network periphery (Kelly et al., 2006), while the study on the discussion network of a fan group found that members tried to engage the troll to understand and change the troll’s behavior rather than ignoring it altogether (Baker, 2001). Typically, the active engagement not only failed to change the trolls’ attitudes or make them comply with community norms but also wasted community members’ time and disrupted on-topic discussions. Thus, instead of improving the community environment, there is consensus that, as compared to other reactions, actively responding to trolls harms the community itself.

Therefore, it is important to understand the kind of community members who would engage with trolls and the conditions under which such engagement happens. Klempka and Stimson (2014) found that people in the older age group are more likely to respond to trolls. Those members were identified with high ego-involvement, and they may experience contradicting norms that opposite the offline communication culture when facing trolls. They are more likely to engage with trolls because they are personally involved in the conversations. But demographics aside, little do we know about the mechanisms leading to different reactions. No extant research has examined the characteristics of the communities and their members in relation to reactions to trolls. Drawing from social identity theory, the current study uses social network analysis to systematically identify people who will reply to trolls and community characteristics that facilitate active reactions to trolls.

When community membership becomes their salient social identity, members will act upon particular social norms and beliefs associated with their membership. Especially when facing threats, where the social being of ingroup members are endangered by an outgroup and their actions, community members with a salient social identity are more likely to protect their ingroup, defend their community, or even attack the outgroup (Reicher et al., 2008). Those behaviors are empowered and supported by their social identity. When facing trolls, who are perceived as threats to the community, members who have a stronger perceived social identity are more likely to respond to trolls.

Social identity with an online community can be constructed through group affiliations, which require a time investment, emotional involvement, and frequent participation (Bergami & Bagozzi, 2000). The group affiliation, in return, motivates an individual to interact with other group members, which may lead to various social networks. In an online community, social network structure plays a functional role in information exchange among its members. It can provide access to potential resources to fulfill one's needs (Zaheer & Bell, 2005) and also maintains the performative acts of social identification online (Pearson, 2009). Thus, social network analysis can be usefully combined with SIT in this study to examine community members' reactions towards trolls. Previous data-driven research on trolls illustrated those reactions to trolling behaviors could be predicted with network features of online communities (Al-garadi et al., 2016; Squicciarini et al., 2015).

Previous research has shown that identification with an online community is related to social tie strength and homophily (Brown et al., 2007). Social identity is maintained by online behaviors within the network structure (Pearson, 2009), such as joining online groups, posting in forums, and interacting with other group members. Ingroup affection and ingroup ties are formed through these processes (Cameron, 2004). The strength of those ties is a combination of the amount of time spent together, the emotional intensity, the intimacy, and the reciprocal services (Granovetter, 1973). As Wang and Fesenmaier (2004) pointed out, interactions among community members are essential to community prosperity that it reflects members' commitment to it. The process of getting to know each other and communicating enables the social influence of a community (Postmes et al., 1999), while interpersonal ties with other group members and the information exchanged through those ties will benefit individuals socially and psychologically, reinforcing the group affiliation and enhance social identity (Cameron, 2004; Millen & Patterson, 2003). In other words, people with a more salient social identity are more likely to get involved in the cohesiveness of the network.

Individually, those who have more ties and are more connected in the network have better access to information and resources, which may help them solidify influence in the community. As they become more important in the community, their behaviors tend to be accepted and trusted (Reicher et al., 2008), leading to positive ingroup distinctiveness and a salient social identification. Thus, community members situated at more central positions in a network may have a stronger perceived social identity attached to the community. Those members may seek the positive distinctiveness of an ingroup, and the behavior they choose may reflect their perceived stability and legitimacy of their social identity. When facing trolls, those community members may be more active or even more aggressive, and they may tend to reply to and argue with trolls to defend the norms of the community. Here, we hypothesize that:

H3.1: Community members who are more central in the community network will be more likely to respond to trolls.

According to Tsai and Men (2013), identification with a community is both the cause and the indicator of members' engagement within the community. The members with a stronger perception of social identity attached to the community are likely to be more actively engaged in the community, leading to stronger ties with other group members that benefit them socially and psychologically. These strong ties formed by active members also increase the attachment to and the social identification with the community. Therefore, active community members are likely to have a strong perception of social identity with the group. When facing trolls, those members tend to engage trolls in discussions.

H3.2: Community members who are more active in the online community will be more likely to respond to trolls.

When most community members hold strong perceptions of social identity, the community is likely to be perceived as a concrete entity, as the community members are highly engaged in normative behaviors, such as posting and commenting. Community members interact with each other actively, which leads to plenty of strong social ties (Sohn, 2009). Therefore, it is likely for a community with a large number of members who have a strong social identity to form a dense social network. Vice versa, community members embedded in a dense network are more likely to have strong ties with other community members, indicating a high psychological and emotional attachment to others and a strong affiliation with the group. They are more willing to contribute to the community without caring too much about their own personal gain but rather care more about the interest of the group. Community members in a dense network are more likely to have a high social identity with the community. When facing trolls, those members tend to protect the shared interest of the group and act defensive towards trolls.

H3.3: Community members in densely connected networks will be more likely to respond to trolls.

The process of identification with a community not only cognitively categorizes individuals but also encourages individuals to learn from observations and exchanges within a community (Peteraf & Shanley, 1997). According to social learning theory, behaviors can be acquired by direct instructions as well as observations and imitations (Bandura & Walters, 1977). In an online community, individuals who identify themselves with the community can learn normative behaviors. Online communities provide an environment in which exposure to normative behaviors, imitation of models, and social reinforcement take place (Akers et al., 1995). Learning occurs when community members adapt their behaviors by reciprocal interaction and observation, and social reinforcement can motivate the learning process (Bandura & Walters, 1977). Through interaction, group members gain information about group norms and values. By observing others' behavior towards ingroup and outgroup, members learn what behavior will benefit the group, elevate self-esteem, and raise ingroup status (Nicholson & Higgins, 2017). Therefore, when a community member observes a series of reactive behaviors towards trolls, they learn to imitate those behaviors, reinforcing them further as they are considered socially desirable or even normative. Thus, we hypothesize:

H3.4: Community members will be more likely to respond to a troll if a) there are more prior responses to trolls in general in the community, and b) there are more previous responses to that specific troll.

Social interaction is a key element in social learning (Hogg, 2016), and the influence of the group is more pronounced if it comes directly from friends rather than strangers. Community members can learn the accepted or normative reactions to trolls by observing their existing social ties. Learning from existing ties is more likely to occur because past interactions promote trust reinforcing the behaviors. An individual is even more likely to adopt a behavior if multiple, rather than one, social ties exhibit the same behavior (Centola, 2010). Therefore,

H3.5: Community members will be more likely to respond to a troll if a) other community members whom they have interacted with responded to trolls in general, and b) other community members whom they have interacted with responded to that specific troll.

Social norms are informal understanding that governs the attitudes and behaviors which characterize a social group and differentiate it from other social groups (Hogg & Reid, 2006). It generates positive distinctiveness so that group members are motivated to behave consistently with the shared understanding to enhance ingroup social identity (Christensen (Christensen et al., 2004). Therefore, individuals with stronger perceptions of social identity may feel good about themselves if their social relationships are congruent with group norms (Wood et al., 1997). People who socially identify themselves with a group are more likely to conform to the group norms. Norms grow more salient as more people observe the same belief or behavior (Kelly et al., 2006). Specifically, how to respond to the trolls might become a commonly held norm among members. For example, community members from a feminist forum reached an agreement about how to appropriately respond to a troll after being attacked by that troll (Herring et al., 2002). Such an agreement can be codified and normalized in the community and would be extended to all trolls in general. Therefore,

H3.6: Community members will be more likely to respond to a troll if a) more community members respond to trolls in general, and b) more community members respond to that specific troll.

Prior research in psychology and organizational communication has shown that people respond differently to positive and negative stimuli and that negative events tend to elicit stronger and quicker emotional, behavioral, and cognitive responses than neutral or positive events (Rozin & Royzman, 2001). Negativity bias is defined as "…a general bias, based on predispositions and experience, to give greater weight to negative entities" (Rozin & Royzman, 2001, p. 296). Taking an evolutionary perspective, negative information likely alerts potential dangers and negative outcomes. Thus, negative entities are perceived more strongly than equivalent positive entities. Negative messages are generally considered to gain more attraction and responses (Knobloch-Westerwick et al., 2020). Therefore, messages with negative sentiment are likely to provoke community members to watch out for potential threats towards their online community, and they would in turn respond more strongly to these potential threats. In addition, negative entities receive more varied responses. For instance, in linguistics, there is more vocabulary used to describe the qualities of negative events (Rozin & Royzman, 2001). Recent research has shown that negative sentiment posts induce more feedback than positive sentiment (Stieglitz & Dang-Xuan, 2013). Therefore,

H3.7: Trolling posts with higher negativity are more likely to provoke responses.

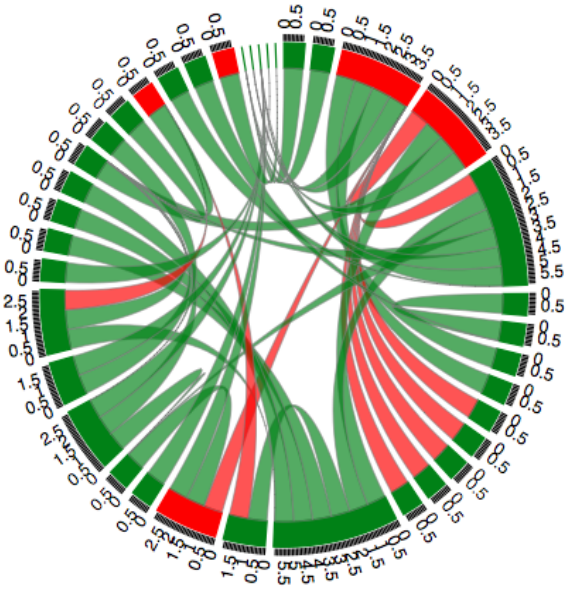
3.2. Methods

3.2.1. Data

Prior research has shown that YouTube video communities are fertile ground for observing collective identities (Halpern & Gibbs, 2013). This study collected data from communities formed around individual videos on YouTube. These communities consist of individuals who are interested in videos on particular topics and further engage in discussions through comments. Data were collected from the top 3 most subscribed political channels (The Young Turks, CNN, and BBC news) in February 2018 with comment section enabled, and top 3 most subscribed comedy channels (CollegeHumor, Annoying Orange, and PowerfulJRE) in February 2018. Each channel had more than 5 million subscribers at the time of data collection. All selected videos are posted in the period from Dec 14, 2017, to Jan 4, 2018. This resulted in 631 videos from 6 channels, with 2200 comments on each video on average. From which we further selected video communities if they satisfy three criteria: 1) the video was between 4 to 10 minutes long so that it provides adequate information for potential discussions; 2) the video had been uploaded at least one month prior to the time of data collection, allowing enough time for interactions via commenting to take place; and 3) there were at least 1000 comments posted (excluding deleted comments). As the average length of the 631 videos was around 7 min, the 4-10 mins interval was selected to ensure that the videos have comparable lengths and content density.[1] The average number of comments from those videos was 2341 (*SD* = 1365), which led us to select 1000 as the threshold (around one standard deviation from the average). The one-month and 1000-comment thresholds were applied to ensure that enough social interactions took place to form community identities and that there was sufficient time and attention for trolls to provoke and receive responses. Twenty-three video communities fitted our criteria, within which the number of unique individuals ranged from 11 to 344 (*M* =160.52, *SD* = 93.12). The number of comments per video ranged from 1081 to 5151 (*M*=1930.46, *SD*=1045.72).

To test the hypotheses, we need to know a community's communication network dynamics before a troll speaks in the community. Hence, the pre-troll community networks were created through YouTube's threaded commenting structure. Nodes represent individuals in each community, and a direct link from individual A to individual B represents a reply from A to B. As deleted comments could not be recovered, any thread involving deleted comments was excluded from the networks. From all 23 communities, 8299 (18.11%) comments were eliminated because they were in threads involving deleted comments (*M*=360.93, *SD*=250.56). For a specific troll A, the pre-troll community network was constructed from all comments before troll A entered into the community. Similarly, pre-troll networks were created for the first five trolls only in each video community with more than five trolls because pre- and post-troll dynamics become less distinct as more trolls infiltrate the community. In total, 874 unique individuals were identified with 3692 directed links from 111 pre-troll networks in 23 video communities. A community of 38 members with the first 5 trolls is illustrated in Figure 3.1.

Figure 3.1. *Interactions among trolls and community members in one YouTube community.*



Note. Each bar represents an individual user in the community. The red bars represent the first five trolls in the community. The number on top of each bar reflects their outdegree. The strip linking two bars represents a tie between those two individuals. The wide end of the strip is the individual who make the comment, and the narrow end is the individual whose post was commented.

Following prior research conducted by Al-garadi et al. (2016), a pre-trained Naive Bayesian classifier was used to identify trolling messages. The Naive Bayesian classification model was built with a vulgarity word list, second-person pronoun, and the sentiment of the comment (Al-garadi et al., 2016; Squicciarini et al., 2015). We collected online bad word lists from banbuilder.com, bannedwordlist.com, noswearing.com, and urbanoalvarez.es, which contain offensive words and cursing words used by native English speakers on social media (Wang et al., 2014). According to previous research, second-person pronouns are highly related to online anti-social behaviors (Squicciarini et al., 2015). We considered it as a criterion to identify trolls. Trolling messages usually contain negative emotions such as anger and anxiety

(Al-garadi et al., 2016; Squicciarini et al., 2015). Therefore, the sentiment was included as the third criterion for troll detection. In addition, 300 comments were randomly selected and coded by a human coder to validate the accuracy of machine classification. Cronbach's alpha of human coding and machine labeling was 0.86. We define a troll as an individual who has posted at least one trolling message in one community. In total, 1399 trolls were identified from all 23 communities (*M*=60.83, Median =51, *SD*=35.53), with a minimum of 1 troll and a maximum of 249 within each community. These trolls posted 2506 messages in all 23 communities (*M*=111.74, Median =96, *SD*=68.93).

3.2.2. Measures

3.2.2.1. Likelihood of Responding to the Troll. For any specific troll, the likelihood of responding to the troll was defined as a binary variable, where 1 indicates the community member responded to the troll in a comment. For 111 trolls in 23 communities, there were 216 individuals who responded to trolls at least once.

3.2.2.2. Degree Centrality. Degree centrality captures the extent to which an individual is integrated into the community. It is measured by the number of people a community member has directly interacted with. A community member who has a high degree centrality is in a central position in the social network. Specifically, indegree is the number of people a community member received replies from (*M*=0.98, *SD*=2.77), and outdegree is the number of people a community member replied to (*M*=0.98, *SD*=0.86).

3.2.2.3. Density. The density of the networks (*M*=0.43, *SD*=0.19) indicates the connectedness of a network. It is measured by the following formula:

$$D = \frac{The\ total\ number\ of\ actual\ links}{The\ total\ number\ of\ potential\ links}$$

As indicated by the formula above, density is reversely correlated with network size (*M*=47.93, *SD*=25.24, *r*=-0.73, *p*< .001). The number of potential connections among community members grows exponentially when the number of community members grows. Therefore, to test the effect of density while controlling for network size, we regressed density on network size and obtained the residual as the measure of network density.

3.2.2.4. Activeness. Activeness is measured by the total number of an individual's comments and replies in a pre-troll network. In the pre-troll network of troll A, community member B's activeness is the total number of B's original comments and replies to other community members (including trolls appearing before troll A). It is an important indicator for engagement (*M*=1.59, *SD*=1.42). As the distribution of activeness is right-skewed, we use logarithmic transformation $x' = log\ (x + 1)$for activeness (*M*=1.25, *SD*=0.52).

3.2.2.5. Prior Responses. To test how people learn from others, the number of responses to troll A before community member B replied to troll A (*M*=0.35, *SD*=1.02) and the number of responses to all trolls in the pre-troll network of troll A before community member B replied to troll A (*M*=2.12, *SD*=2.69) were calculated. Similar to the measurement of activeness, both the number of prior responses to one specific troll and the number of prior responses to trolls, in general, were right skewed with a high frequency of 0 value. Logarithmic transformation was applied to both measurements (See Table 3.1).

Table 3.1. *Descriptive Statistics of Main Variables (N= 3692).*

Variable	*M*	*SD*
Individual level		
Indegree	0.98	2.77
Outdegree	0.98	0.86
Activeness	1.59	1.42
Activeness(log)	1.25	0.52
Number of prior responses to the troll	0.35	1.02
Number of prior responses to all trolls	2.12	2.69

Responses from existing ties to the troll	0.001	0.02
Responses from existing ties to all trolls	0.025	0.21
Response (binary variable)	2.38%	
Troll level		
Number of responses to the troll	0.61	1.60
Density	0.03	0.0
Negativity (%)	0.83	0.16
Number of individuals in the network	47.93	25.2
Community level		
Number of responses in community	3.38	3.18
Number of responses in community (log)	1.66	1.23

3.2.2.6. Total Responses. The total responses to the troll were defined as the total number of responses to troll A in the community (*M*=0.61, *SD*=1.60). It is included in the analysis to test how social norms in a community affect people's reactions to trolls. The total number of responses to the troll is found right skewed with 0 value, so a logarithmic transformation is applied. In addition, the total number of responses to all trolls in the community (*M*=3.38, *SD*=3.18) was calculated by counting the total number of replies from community members to any troll. Again, a logarithmic transformation is applied to address the right-skewed distribution.

3.2.2.7. Existing Ties Who Responded to Trolls. To test whether people learn more from those they have communicated with, community members' existing ties to other members who replied to trolls have been calculated. For each community member, responses were counted when they replied to those who respond to troll A in the pre-troll-A network (*M*=0.001, *SD*=0.02). Specifically, responses were counted when community member B replied to community member C if community member C has replied to troll A. Responses were also counted when they replied to those who respond to any troll in pre-troll-A networks in the community (*M*=0.025, *SD*=0.21).

3.2.2.8. Negativity. Sentiment analysis was applied to get negative sentiment (*M*=0.82, *SD*=0.16). The score is obtained from text mining tools MonkeyLearn (Kaur & Chopra, 2016; MonkeyLearn, 2020). MonkeyLearn is an online text mining service for classifying text information for specific needs. Research has shown that it performs well in detecting negative sentiments with high accuracy (Basmmi et al., 2020) using neural network models. The negative score is the probability of negativity of the text. It ranges from 0 to 1. A higher score indicates a more intense negative sentiment.

3.2.3. Analysis

The dataset was arranged at the individual-troll-community level, with each row representing a community member's responses (or lack thereof) and network measures within a specific pre-troll network in the community. In other words, each community member was nested within each specific troll, which was further nested within each community. Due to the multilevel data structure, we used mixed-effects regression models to ensure that variations between trolls and communities were properly accounted for. Thus, to test Hypotheses 1-7, mixed-effect logistic regressions were applied to predict the likelihood of each community member responding to trolls. The dependent variable was whether a community member responded to a specific troll (1= yes). The independent variables included the community member's centrality, activeness, network size, density residuals, the negativity of the troll's post, community's total response as well as prior response to the specific troll and to all trolls in general, the community member's existing ties who replied to the specific troll and those who replied to any trolls. The models include nominal variables for each troll and each community as random effects.[2] The results are presented as Model 1 and Model 2 in Table 3.2.

Table 3.2. *Logistic Regression Models Predicting the Likelihood of Response to Troll*s *(N=3692).*

	Model 1				Model 2			
	B		*SE*	*OR*	*B*		*SE*	*OR*
Intercept	-6.677	***	1.032	0.001	-8.040	***	1.040	0.000
Outdegree	0.518	**	0.173	1.679	0.515	**	0.159	1.673
Indegree	-0.605	**	0.197	0.546	-0.648	**	0.212	0.523
Total response to the troll	1.921	***	0.200	6.828				
Prior response to the troll	-0.310	.	0.169	0.733				
Responses from existing ties to the troll	-14.632		362.039	0.000				
Total response to all trolls					1.048	***	0.190	2.851
Prior response to all trolls					0.332	**	0.116	1.394
Responses from existing ties to all trolls					0.514		0.334	1.671
Negativity	1.079		0.971	2.940	2.797	**	0.858	16.401
Activeness	0.451		0.320	1.570	0.428	.	0.260	1.534
Network size	-0.038	**	0.012	0.963	-0.064	***	0.011	0.938
Density residual	0.190	*	0.078	1.209	0.084		0.063	1.088
Model AIC	423.7				610.2			
Log likelihood	-199.9				-293.1			

Note. '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

3.3. Results

The first hypothesis investigated whether a community member's centrality in a pre-troll network of troll A is associated with the probability of the community member responding to troll A. Controlling for network size, indegree was negatively related to responding to the troll (odds ratio=0.55, *p*<0.01) and outdegree was positively related to responding to the troll (odds ratio=1.68, *p*<0.01), both dimensions of degree centrality showed a significance in predicting the possibility of responding to the troll, supporting H1.

H2 tested whether a community member's activeness in the pre-troll network of troll A is associated with their responding to troll A. We found no significant relationship between community members' activeness in the pre-troll network and the likelihood of responding to a troll (odds ratio=1.57, *p*= .16). Therefore, H2 is not supported.

H3 predicted that a pre-troll network's density is positively related to the likelihood of a community member responding to a specific troll in that network. Controlling for network size, pre-troll network density was positively associated with responding to the troll (odds ratio=1.21, p<0.05), supporting H3.

H4 investigated whether a community member's reaction to a specific troll is associated with a) other community members' prior responses to the specific troll and b) their prior responses to all trolls in general. Controlling for network size, a prior response from other community members to a troll in the pre-troll network is not significantly associated with community members' response to the troll (odds ratio=0.73, *p*=0.067), so H4a is not supported. And prior responses to all trolls in a pre-troll network are positively associated with the community member's response to a troll (odds ratio=1.39, *p*<.01), supporting H4b.

H5 tested whether the probability of community members' responding to a troll would be affected by their existing ties who had responded to the specific troll (H5a) or who had responded to all trolls in general (H5b). We found no significant evidence supporting H5a (odds ratio=0, *p*=0.97) and H5b (odds ratio=1.67, *p*=0.12).

H6 tested whether a community member's response to a specific troll was affected by all other community members' responses to the specific troll (H6a) or their responses to trolls in general (H6b). Both H6a and H6b were supported, with the total response from other community members to a troll (odds ratio=6.83, *p*<.001) and total response from other community members to all trolls (odds ratio=2.85, *p*<.001) positively related to a community member's likelihood to respond to a troll.

H7 tested whether negativity of comments was associated with a community member's response to a troll. Results showed that negativity was positively related to the likelihood of response from a community member to a troll (odds ratio=16.40, *p*<.001).

3.4. Discussion

As a widely observed anti-social behavior, trolling disrupts on-topic discussions and irritates community members by inserting misinformation, misleading judgments, and personal attacks into the fabrics of online discussion. When members actively engage with trolls instead of shunning them all together, it wreaks havoc on the well-being of online communities, leading to the decline of various community functions and the retreating of legitimate members. By analyzing the social dynamics of popular video communities on YouTube before and after trolls, we focused on the likelihood and conditions in which community members would respond to trolls. We found that the valence of the trolling message, characteristics of individual users, as well as the patterns of past engagement with trolls from other community members all played a role in shaping how an individual would react to trolls. In other words, well-connected members who are situated in densely connected communities with a prior pattern of responding to trolls are more likely to follow suit, especially when the trolling messages convey negative sentiment.

First, while most previous research focused on the structural position of trolls in the network (Kelly et al., 2006), this study makes a contribution to trolling literature by demonstrating that the community's existing structure and each member's position in the network also mattered in member's response to trolls. Drawing from social identity theory, community members who have higher centrality, especially outdegree centrality, are more likely to have a stronger social identity and higher social status in a community. Their behaviors are governed by the sense of ingroup distinctiveness. They are more willing to defend their group norms. The density of a pre-troll network is also positively associated with members' responses to trolls. In a dense network, community members are better connected with each other, and shared information and ideas are more likely to reach most community members in a short time, ensuring salient group values and norms. Community members hold a stronger ingroup identity and are more likely to respond to those who challenge their values and norms.

The second notable finding is that individual members' reaction to trolls also depends on the community's collective responses (prior as well as cumulative), which can be explained by social norms and social learning processes. In other words, just as toxic behaviors themselves are contagious in online communities (Shen et al., 2020), the reactions to toxic behaviors also spread from member to member. A community's prior collective responses to trolls, in general, are positively related to an individual member's likelihood to follow suit, yet prior collective responses to the specific troll do not have the same effect. One explanation is that perceived community norms are only formed -- and to be imitated -- when such response patterns are repeated against multiple offenders, whereas collective responses to one specific troll do not reach the threshold to get normalized. Further, our hypothesis that community members are more likely to follow their existing ties in responding to trolls was not supported. One possible reason is that in our dataset, pre-troll networks tend to be sparse, and thus there are not enough data to test the effect of existing ties.

We found that post sentiment was also associated with a community member's likelihood of responding to trolls, confirming negativity bias. The more negative a trolling post is, the more likely it will elicit responses. This finding resonates with previous research (Rozin & Royzman, 2001) as negativity triggers more feedback. However, the activeness of the community member was not associated with the probability of responding to a troll. One potential explanation is that activeness, as measured by frequency of posting, does not necessarily equal a strong social identity or high ingroup status. This was manifested in the low correlation (r =0.38) between degree and activeness.

An important practical implication can be gleaned from the study for designers and managers of online communities alike. It is widely recognized that trolls are detrimental to the health of online communities, yet most past research and practical solutions have focused on the identification of trolls and proper sanction mechanisms to keep them at bay. This study hints at a different strategy altogether, one that recognizes the importance of establishing and reinforcing group norms for appropriate reactions to trolls (i.e., do not feed them) before trolls emerge. It shifts the focus from *re*-acting to the attacker to *pro*-actively planning out the community's defense. Our findings suggest that it is most effective to advise all members, especially well-connected ones, against responding to trolls before trolls appear. When a few members inadvertently respond to trolls, they set up examples for other members to follow, as response begets response. The conversation could devolve quickly and significantly derail normal community functions. Therefore, efforts aiming at educating the community about troll identification and prevention pay far more dividends than reactive strategies.

3.5. Limitation and Future Research

Our study has a few limitations. First and foremost, the establishment of communities from YouTube videos might be not as solid as expected. We assumed that videos could generate online communities and form a sense of loyalty and identification from users who participant in the discussions related to the videos. In other words, we assumed that users who has the motivation to discuss about the videos are those who already feel attached to the topics or ideas presented in the videos, so they are likely to spontaneously form a community with a network structure. And the sense of attachment to the topics will transform to the social identity with the community. However, the establishment of a community usually takes time, which is less likely to form over the time of a video clip. Thus, the individual and collective reactions to trolls may not be the results of social identification towards communities, but are the consequences of other confounding variables, such as users' political affiliations, personality, and Internet use history. For instance, most videos in our dataset are politically oriented and may be perceived as liberal leaning, which may trigger users' political believes, and promote them to react to comments that they disagree with. In addition, previous research has shown that certain personality traits may positively related to trolling behaviors and victimization, such as social extraversion and depression are related to trolling behaviors and a sense of inferiority is related to victimization (Hong & Cheng, 2018). Users who reply to trolls may also have personality traits that associated with the behavior. Similarly, user's pervious internet using, like reading topic related articles, debating with others, and playing video games, may also affect how they react to trolls. Established and structured online communities should be taken into consideration for future research, and other methods to assess users' political affiliation, personality, and internet use history can be applied in the future.

The second limitation is that mentioned above, most videos in our dataset are politically oriented and may be perceived as liberal leaning, which may be prone to trolling behaviors. Future research can examine more diverse content categories and on more platforms. How political leaning will affect people's tolerance and reactions towards trolls may also be a good direction for future research. Third, the study is limited by the characteristics of the dataset. It is cross-sectional data that cannot offer the full view of responding dynamics. Besides, the networks have relatively low density, with low centrality and modularity, and skewed distributions. In addition, the captured networks are not highly active networks, so the total response rate is low. Future research would benefit from including more active and more varied network datasets. Fourth, as there are many different operationalizations of trolls, our specific operationalization may not fit other types of communities. Further research could adopt and validate different operational definitions of trolling, which would bring a better understanding of online anti-social behavior. Besides, given our findings, future studies can further zoom into a few insignificant variables as well as variables we failed to include. For example, this study only examined a limited number of communities and their members' interactions within these communities. The prior interaction history of community members outside of these communities, against the backdrop of YouTube as a giant primordial social network, may also offer interesting insights with regard to social identity and social norms. In addition to negative sentiment, more discrete sentiments can be studied to understand the affective interactions with trolls in online communities.

3.6. Conclusion

Trolls provoke emotional and behavioral responses from ordinary internet users, disrupt the normal functioning of online communities, and erode their social fabric. To mitigate such harms, researchers and practitioners need to understand not only the trolls themselves but also the conditions in which their intended audience -- the legitimate community members -- receive and respond to their provocations. Analyzing conversations on YouTube communities, this study demonstrates that centrality, community network density, and negative sentiment in trolling messages all contributed to a member's likelihood of responding to trolls. These findings highlight the importance of group norms and regulations in developing trolling prevention and intervention mechanisms. Identifying and banning trolls is only half the battle, actively fostering group norms such as “don’t feed the troll,” especially among well-connected members, is also an effective way to safeguard the well-being of online communities.

Chapter 4. Will You Become the Next Troll? A Computational Mechanics Approach to the Contagion of Trolling Behavior

As human interactions in digital environments increasingly became an important part of our daily activity, the concerns about misconduct in online platforms increased. According to a national survey in 2020 (Pew Research Center, 2021b), 41% of people in the U.S. have experienced any form of online deviant behaviors, among those who have experienced deviant behaviors, 45% attributed it to their political views, and 75% encountered on social media. One of the deviant behaviors that draw the most attention is trolling behavior. Trolling is considered a deviant behavior that diffuses misinformation, provokes emotional responses, or disrupts on-topic discussions (Shin, 2008). The motivations of the behavior vary from simple attention attractions (Herring et al., 2002) to well-calculated manipulations of public opinions (Engelin & De Silva, 2016). In addition, the effects and consequences of the behavior are also varied across audiences and contexts, leading to significant problems that affect individuals and online communities.

Much of the work on trolling behavior has focused on its language aspect, investigating the identification and antecedents of trolling behaviors (Coles & West, 2016; Engelin & De Silva, 2016; Hardaker, 2010; Lökk & Hallman, 2016). Scholars believed that capturing the feature of messages would solve the puzzle of "what," "how," and "why." The premise is that if someone talks like a troll, it is a troll. While as a complex online deviant behavior, trolling is much more than verbal abuse. It is a behavior to attract responses and gain influence by making time-wasting comments on controversial topics, using provocative language and strategies, such as referring to a person, repeatedly commenting, and using misinformation. Therefore, our premise is that if someone behaves like a troll, it is a troll.

Per definition, behavior is a dynamic process. However, traditionally, Communication as a field has not paid much attention to dynamic processes that unfold in time (Poole, 2007). In this project, we aim to bring the behavioral perspective to deconstruct the communication dynamics of trolling. With traceable digital footprints from Reddit in 2016 and 2017 and autoregressive model-based investigation into human-human interactions, we represent important new opportunities to examine whether and how trolling behavior is contagious on a large scale. With the help of the unique, minimally complex, maximally predictive model of the dynamic, the so-called predictive state model (Shalizi & Crutchfield, 2001) answers the following questions: (1) Is the trolling behavior self-activated or situational? (2) How much is trolling behavior contagious? (3) Is there any hidden pattern of trolling behavior?

Our models show a complex hidden pattern guiding individuals' trolling behaviors. Engaging in trolling behaviors can be self-motivated, social-motivated, and self-social-together-motivated. In addition, individuals are more likely to adopt trolling behaviors from people they encounter rather than becoming trolls by themselves. Counter-intuitively, individuals who are more active in online space are not necessarily more likely to be influenced to engage in trolling behaviors.

4.1. Trolling as a Dynamic Process

Scholars have not reached a conclusion on a universal definition of trolling; rather, previous literature defined trolls in various ways. Some focused on the deceptive nature of trolling messages (Breeze, 2012; Donath, 2002; Engelin & De Silva, 2016), claiming that trolls use fake persona and misleading messages to impose opinions on others, damaging trust in online communities, and find valuable information from others. Others argued that provocation is the main feature of trolling (Bergstrom, 2011; Hardaker, 2010; Merritt, 2012; Nevin, 2015), where trolls use erroneous or inflammatory information to trigger negative reactions from ordinary internet users. In addition, trolls are defined differently context-wise. For example, in gaming platforms, trolling is a clear concept associated with griefing culture (Cook, 2019) and treated as a strategy for competition. While in the domain of political discussion, trolls are generally considered ideological devices that express extreme political opinions and mislead their audience (Fichman & Sanfilippo, 2016). In this study, we approach from a comprehensive perspective that trolling is a behavior with the aggressive characteristics and misleading nature, echoing Cheng et al. (2017) inclusive definition where trolling is behavior “falls outside the acceptable bounds” (p. 1217) and Herring et al. (2002) definition which focused on both characteristics and outcome of the behavior.

4.1.1. Communication as a Dynamic Process

Traditionally, trolling behaviors were studied from the textual perspective, focusing on the semantics and sentiment of the text (Clarke, 2019; Komaç & Çağıltay, 2019), but communication behaviors have been studied as dynamical systems long before from an information theoretical perspective (Cappella, 1979; Ellis & Fisher, 1975; Fisher & Drecksel, 1983). From such a perspective, information is viewed as the resolution of uncertainty (Thomas & Joy, 2006), providing a method to quantify the predictability of the dynamic systems of communication. As a complex dynamical system, an individual’s communication behaviors are considered discrete components or constituents that are part of a collective, in which each component or constituent will affect one the other to create a pattern, an order, or a structure at the level of the collective that cannot be observed on the individual level (Fogel, 2006). Furthermore, communication behaviors, in general, involve more than one individual, where complex dynamics happen within inter-individual relationships. the dynamic of communication behaviors is likely to involve other individuals, contingent on the ongoing and simultaneous flow of communicative actions from both individuals involved, creating patterns, orders, and structures. In such a co-dynamic system, individuals' behaviors are dynamically altered by their social partners.

4.1.2. Trolling as a Communicative Dynamic Process

Previous research has argued that people who engage in trolling behaviors have unique traits and personalities (Buckels et al., 2014; Hardaker, 2010), while other research suggests that ordinary people can be triggered into engaging in trolling behaviors by discussion content and sentiments (Cheng et al., 2017). On one hand, engaging in trolling behaviors could be a well-practiced behavior recuring in constant contexts, reflecting individuals' personalities (e.g., sadism, Buckels, et al., 2014), biological traits (e.g., low baseline arousal,), or operational mechanism developed by acculturation to internet culture. On the other hand, when individuals read, comment on others' posts, and reply to others' comments, they may be influenced by the contextual environment or even by the one they interacted with. The provocative nature of trolling behaviors may trigger mood swings. Individuals may experience negative emotions and further engage in trolling behaviors to outlet emotions and get revenge on the previous troll. Experiencing previous trolling behaviors may raise the likelihood of individual engaging in future trolling behaviors. Vice versa, it is also possible that angry individuals who post trolling messages will calm down after reading some non-trolling comments.

We adapt the approach of information theory (Thomas & Joy, 2006), where the swings between trolling behaviors and non-trolling behaviors, thus, can be framed as a dynamic system of information processing. Follow the lead of Shannon (Shannon, 1951), the trolling dynamic system can be chronologically viewed as temporal sequence. Therefore, the perspective shifts our scholarly attention from individual trolling messages to a temporal sequence combining trolling and non-trolling behaviors. Whether and how individuals engage in trolling behaviors can be understood as related to the co-influence between trolling and non-trolling behaviors, and mathematical frameworks can be applied to describe the complexity of the dynamic. Here, we define complexity as a measurement of the amount of patterns that can be identified in the dynamical process (Crutchfield, 1994). In other words, the mathematical frameworks can uncover dynamical characteristics of the temporal sequence of trolling and non-trolling behaviors, identifying how trolling and non-trolling behaviors interact with each other.

According to previous research on behavior pattern (Darmon, 2015), such temporal sequence can be analyzed as an autoregression that their own past behavior influences their future behavior. While another possibility is that users are affected by contextual environment, where they react to others instead of posting independently. If so, the dynamical characteristics of the users' temporal sequence should reflect on such relativeness, where a user's temporal sequence will be associated with another user's temporal sequence. Furthermore, users' behavior may be much more complex with multiple influences that their own history and the contextual environment can have an impact together on their future behaviors. Thus, we first ask,

RQ4.1: Is trolling behavior innate or situational? Or both?

Information theory also provide a new approach to social influence, diffusion, and contagion models. Traditionally, social influence is studied on a macro level, as a fundamental force that drives the formation and propagation of psychological states (e.g., emotions, Hatfield, et al., 1993), opinions and attitudes (Wood, 2000), and behaviors (Wheeler, 1966). Research focused on using contagion, social influence, and social learning models to analyze whether and how people massively reach or adapt certain psychological states, attitudes, or behaviors (Herrando & Constantinides, 2021), while information theoretical perspective can zoom in to individual behavioral dynamics, providing an opportunity to investigate direct influence on individual level.

Previous research has showed that human tend to align the emotional states they receive during communication interactions instinctively (Ekman et al., 1983). Emotional contagion is reflected in such a facial, vocal, or postural alignment, as well as similar neurophysiological and neurological reactions (Hatfield et al., 1993). Moreover, emotional contagion is also showed in computer-mediated communication (Herrando & Constantinides, 2021). For instance, in 2014, the controversial Facebook emotion experiment (Kramer et al., 2014) demonstrated in the study and from the Facebook users' responses to the study that emotions can be increasingly intense during the spread on social media, where individuals can experience the similar emotions, resulting in similar emotion expression. As a response to emotional contagion, individuals will show behavioral synchrony (Hatfield et al., 1993), leading to a behavioral contagion. In our case, the contagious negative emotions may increase the probability of individuals engaging in trolling behaviors. With the intensified emotions, trolling behaviors may persist across those affected people to spread further.

In addition, previous research also investigated contagion of deviant behaviors with the mechanisms of generalized reciprocity and third-party influence. Generalized reciprocity refers to victim recouping by "paying it forward" (Tsvetkova & Macy, 2015), while third-party influence is the mimicry after observing the behavior of others. It echoes the hypothetical normalization theory (Beres et al., 2021; Hilvert-Bruce & Neill, 2020; Huesmann & Eron, 1984) that deviant behaviors may not be taken personally by victims and observers, rather attributed to a societal pattern instead. Approaching from the dynamic system perspective, such phenomenon can be described by dynamical characteristics of the interactions among individuals. The temporal sequence of an individual's trolling behavior may show a pattern that the probability of trolling behavior is increased with trolling behavior from contextual sequences. As such, trolling behavior could be contagious and could spread on social networks in the process of interaction and discussion. Thus, we ask,

RQ4.2: Is trolling behavior contagious?

4.1.3. Predictive State Models

The computational mechanics' approach allows us to discern the hidden mechanism about how the information is stored, processed, and transformed over time, constructing a unique model that maximizes the predicted power but minimizes the complexity for the discrete-state, discrete-time stochastic process to describe the dynamical characteristics of temporal sequences of human behaviors. In other words, it allows us to detect behavioral patterns, look at the mathematical architecture of behavior's temporal sequences, and decompose the procedure for generating sequences.

Specifically, the two predictive state models we employed are unifilar hidden Markov models, one with single process and the other with an input-output process. Its hidden states consist of "a set of histories, all of which lead to the same set of futures. It's a simple dictum: Do not distinguish histories that lead to the same predictions of the future" (Crutchfield, 2017, p. 2). Mathematically, the predictive statistic of the past for predicting the future of a conditionally stationary stochastic process is a minimal sufficient statistic for prediction. Thus, for each possible predictive distribution, we could find a class of pasts that induce this predictive distribution, and we can find a statistic ϵ that can map a past into an equivalence class for that past. Here equivalent is defined as “if two pasts result in statistically equivalent future, they are equivalent” (Darmon, 2015, p. 16). The states of a unifiliar hidden Markov model represent a partition for all pasts based on the conditional futures they induce.

Simply put, what we observe in a temporal sequence is a representation of some hidden states, which can be considered as a piece of memory containing what the system will do when receive an input. Usually, those states are not the individual events we can observe but associated with a combination of events. In our case, the hidden states of trolling behaviors can be psychological states, emotional states or other possible states that we may not be able to name without further investigation, for example, the states of anger or not. When individuals are angry, the likelihood of trolling may increase, but it does not guarantee that individuals will engage in trolling behaviors. The angry state does not distinctly map to trolling behaviors. We can observe the trolling behaviors but not the hidden angry state.

The first model is a univariate predictive state model. It is an autoregressive model that is called epsilon machine (Crutchfield & Young, 1989). It quantifies the minimal size, optimally predictive behavior of an individual based on their past. In other words, it assumes that an individual’s future behavior is only influenced by their own past behaviors, so we also call it self-driven model. In self-driven models, the probability of whether an individual troll or not in the future time is determined by whether they troll in time instants before, that is to say, the probability of whether an individual trolls or not in the future time instant t is determined by whether they are trolls in time instants before t.

Practically, at any given time instant t, an individual either troll or not, which can be denoted by X_t. There are only two statuses for trolling behavior $X_t,\ X_t = 1\ or\ 0$, where 1 is assigned to X_t, when the individual’s behavior is identified as trolling behavior, and 0 otherwise.

We assume that the immediate past X_{t-1} is the past behavior that can influence the future behavior, then the probability of the individual has a behavior x_t at time t on the condition of that individual has the behavior x_{t-1} at time $t-1$ is:

$$P(X_t = x_t | X_{t-1} = x_{t-1})$$

The second model expands the autoregressive logic to include a second parallel time series that is considered to influence the ongoing dynamic. Following the modeling approach of computational mechanics, if the epsilon-machine is a finite state machine that computes the individual's future behavior based on its past behavior, we now consider an input-output transducer that computes the individual's future behaviors based on its past and the influence of an external source (Sipser, 1996). Such input-output predictive state models have been called epsilon-transducers (Barnett & Crutchfield, 2015; Darmon, 2015). In our case, we refer to it as the social-driven model. It assumes that an individual's future behavior is influenced by their past behavior and the past behavior of people they interacted with. It aims at capturing social influence and contagion. Thus, additionally to the individual's behavior X_t, we introduce a new variable of social inputs Y_t, and in our case, Y_t indicates the parent comment.

$$Y_t = \begin{cases} 1, & if\ the\ social\ input\ is\ trolling \\ 0, & otherwise \end{cases}$$

Then we have:

$$P\ (X_t = x_t | X_{t-1} = x_{t-1}, Y_{t-1} = y_{t-1}).$$

As in the self-driven model, we derive the unique model with minimal complexity and maximal prediction power for an individual's behavior. The logic of the resulting hidden Markov model is the same as the self-drive model, just that we now have input and output symbols on the transitions (just in ordinary transducers from computational theory). In addition, equivalence classes are identified over joint self and social pasts, and a mapping partition from the current joint past to its equivalence class is detected.

We thus want to trace the computational landscape of trolling behaviors in online environments by constructing the finitary models present individuals' trolling behaviors. The hidden states of the models are important to understand the mechanism of trolling and what motivate individuals to engage in such behavior. Hence, we ask:

RQ4.3: Is there any hidden pattern of trolling behavior?

4.2. Methods

4.2.1. Data Collection

The stochastic models were developed from Reddit comment history from Google BigQuery[8]. Reddit is the ninth most visited website globally (Top, 2018) with more than 330 million users (Alexa, 2019). Reddit has subcommunities designated for specific topics called sub-reddits, where people can post, comment, and react to others' posts and comments. Individuals can subscribe to any sub-reddits to enjoy different content and discussions (while some sub-reddits may have restricted rules for admitting subscribers). One thing that worth noting is that unlike Facebook or Twitter, Reddit encourages users to stay anonymous. It protects users' identities and does not require any identity verification (Reddit, 2021). This policy on one hand supports the freedom of expression (Reddit, 2021), but on the other hand it also fuel the uninhibited trolling, toxicity, or hate speech on the platform. Even with Reddit's anti-harassment policy, we observed trolling behaviors on the platform.

[8]. https://pushshift.io/using-bigquery-with-reddit-data/

We chose to work with the time period of 2016 to 2017, when there was a strong social push for political across-cutting discussion. We identified the 1000 most active accounts on Reddit based on the number of comments they posted during the time period. The comments from those active accounts and the comments they directly replied to, which we considered parent comments, from 2016 to 2017 were retrieved from Google BigQuery. If a comment replied to a post, we also retrieved the post as the parent comment. We obtained a dataset with 17,495,543 comments from 1000 accounts with all their parent comments. Then, one human annotator worked with sample comments from each account to identify Automods, the Reddit's automatic tools to reinforce sub-reddit rules. The Automods were excluded from our dataset, leaving 826 users with a total of 13,456,759 comments, with a maximum of 89,763 comments and a minimum of 9,642 comments (*M* = 16,174.89, *SD* = 9477.37).

4.2.2. Data Annotation and Classification

Seven undergraduate students worked as human annotators for the categorization of troll messages. The target user comments and the parent comments were labeled as "trolling message" and "non-trolling message," respectively, 1 and 0. A coding manual was developed by the annotators and based on Herring et al. (2002) definition. Trolling was identified as comments that included 1) using emotionally provocative language, 2) referring to a person or a group, 3) containing incorrect information, 4) unexpected or unrelated messages. Annotators went through 5 rounds of training and discrepancy solving, reaching inter-coder reliability in a Fleiss's kappa of 0.656. They then moved on to independent annotation. Including the annotation training sets, they annotated 15,558 comments for supervised model training. During the coding process, annotators also encountered deleted and removed comments. Unlike previous research (Sun & Shen, 2021; Sun et al., 2021), where deleted and removed comments were excluded from analysis, in this study, we kept those comments. As we analyzed the individuals' behaviors as time series, excluding time instants from a time series will result in an incomplete time series and misrepresent the mathematical structures, so keeping the "missing value" in the analysis is essential for model building. Moreover, deleted comments were taken down by users themselves that they might regret what had been said, the comments have been downvoted, or the comments received no replies. While removed comments were taken down by platform moderators, usually because of a violation to Reddit or sub-reddit policy. That is to say, removed comments are more likely to be trolling messages but deleted comments may not be necessarily related to trolling behaviors. We coded the deleted comments as non-trolling messages and removed comments as trolling messages.

With the large dataset, we adapted Davidson et al. (2020) incivility BURT model (See Chapter 2) for automatically identifying trolling. Even though the model was not built for the purpose of trolling identification, it has its own advantage when adapting for our dataset. The model firstly used BURT technic to pre-train for domain adaptation on 3 million unlabeled Reddit comments using a masked language modeling objective, which helped capture the Reddit culture. Instead of using incivility annotated data, the model then was fine-tuned with four epochs on 10,000 from the human-annotated trolling comments, with 10% set aside for model validation. The rest 5,558 annotated comments were then used for model testing. The model reached a final *F-1* score of 0.76.

4.2.3. Predictive State Model Building

To build predictive state models, we explicitly focused on the predictive representation of the observed behaviors. As stated above, we follow the practice in behavioral science that reduces the communication flow into a probabilistic sequence. Because the self-driven and social-driven models assume that users' online behavior can be modeled as a conditionally stationary stochastic process, the distribution over futures is independent of the time index conditional on the observed past. To approximate the assumption, we considered each comment as a unique incident happening at a temporal point, and we then had a series of discrete incidents as a temporal sequence of trolling behaviors for each Reddit active user for a two-year time period. Similarly, we also generated a two-year temporal sequence of parent trolling behaviors for each Reddit active user from the parent comments. It is important to notice that our analysis did not consider time stamps. Though we acknowledge that for individual users, a lot of things may happen between they post two comments, and the time intervals may also vary, we have a long time period that are likely to even out the day-to-day differences. In addition, the parallel comparison between self-driven model and social driven model focused on the difference on predictive power introduced by social input, thus, we do not focus on the variance of reaction time.

For both the $\epsilon - machine$ and $\epsilon - transducer$ models, we used the Causal State Splitting Reconstruction (CSSR) algorithm[9] (Darmon, 2015) to infer the models from individual users' time series. The process could be visualized as using a sliding-window with a certain length L to move across the temporal sequence and record the frequencies of different combinations of subsequences. The CSSR algorithm works in two phases: In the first phase, it determines a set of weakly prescient states, and in the second phase, it removes transients and splits the causal states. In this algorithm, L_{max}, the maximum history length which is the length of sliding-window, is used for determining the candidate causal states in the first phase, and size α is the control for the probability of splitting causal states that indirectly controls the total

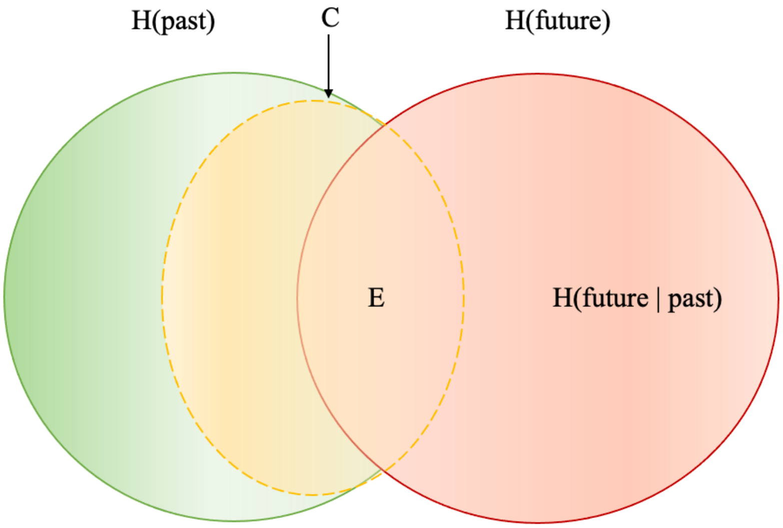
[9] https://github.com/sunyqs/transCSSR

number of the causal states detected by the model in phase two. To achieve an L_{max} with minimum log loss, a train/test split method was used, where half of the data was used to estimate the $\epsilon - machine$ with a specific L_{max}, and the other half was used to determine how well the estimated $\epsilon - machine$ can describe the data. Then the optimized L_{max} was used for machine estimation. Considering that the majority of the comments are not replying to the same parent post, we distinguish among predictive states if they have a similarity less than α = 0.001.

4.2.4. Measurements

Specifically, we calculated three complementary information theoretic measures as we constructed predictive state models, namely, predictable information E, predictive complexity C, and remaining uncertainty h, illustrated in figure 4.1(Crutchfield et al., 2009). The Venn diagram presents past and future information and the communication from the past to the future in terms of entropy H. In information theory, entropy measures the uncertainty of an event based on its probability distribution. Together with the number of states and the maximum history length L_{max}, these measures were used to inform our hypotheses.

Figure 4.1. *Venn diagram of stationary stochastic processes.*



Note. The green circle represents the information needed for past events, and the red circle represent the information needed for future events. As past events can inform what will happen in the future, the overlapped area of the circles represent the mutual information that the past transmits to the future, as predictable information E. The orange dash ellipse represents the predictive complexity C. The area remained in red circle after subtracting the area of E is the conditional entropy $H(future|\ past)$, and the entropy rate h is per symbol rate of $H(future|\ past)$.

4.2.4.1. Predictable Information. The previous literature named predictable information E differently as "effective measure complexity" (Grassberger, 1986), "excess entropy" (Crutchfield & Feldman, 1997), or "predictive information"(Bialek et al., 2001). It refers to the mutual information (Thomas & Joy, 2006) between the past and the future, which is the amount of information that the past communicate to the future. In the Venn diagram (Figure 4.1), it is presented as the overlapping area of the green and red circle. The higher the predictable information is, the more subsequences from the past are used to predict the future.

4.2.4.2. Predictive Complexity. The predictive complexity C quantifies the minimum amount of information that is required for a process to communicate all predictable information (E) from the past to the future (Crutchfield et al., 2009). To optimally predict a process, C is the amount of stored information needed. As the both the $\epsilon - machine$ and $\epsilon - transducer$ models are the unifilar hidden Markov models of the dynamic with the minimum size but maximum predictive power, predictive complexity measures the size of the hidden Markov models, which is showed as orange dash ellipse in Figure 4.1. If we consider E as the capacity of effective information transmission of a communication process, then C is the sophistication of it. The larger C is, the more complex the process is.

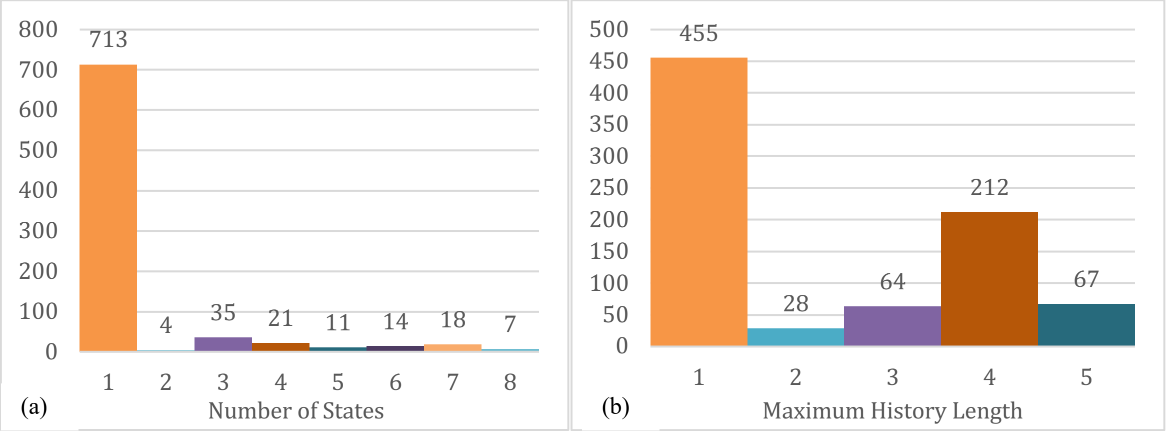
4.2.4.3. Remaining Uncertainty. Showing in Figure 4.1. as the remaining area of the red circle subtracting the overlapped area with green circle, the amount of uncertainty left about the future after using the past to predict the future is the conditional entropy $H(future|\, past)$. It quantifies the information needed for predicting the future besides what can be informed from the past, possibly including information from the outside of the system, channel noise, or measurement error. And we use per symbol rate h to indicate the remaining uncertainty, which is the conditional entropy scaled with the window length L. A larger h means a higher probability of prediction error, in other words, the more remaining uncertainty indicates the larger prediction error, and the more unpredictable future. In addition, in our predictive model, as we stated in previous section, the window length is used as maximum history length L_{max}, the remaining uncertainty rate h is calculated as $H(future|\, past)$ scaled with L_{max}.

4.3. Results

4.3.1. Self-driven Models

We first explored $\epsilon - machine$ architectures across the users. Among 826 users we tested, 713 (86.32%) users only have one causal state, followed by 35 (4.27%) users with three causal states and 21 (2.56%) users with four states. The largest number of states detected from our dataset is eight. The full distribution of states is shown in Figure 4.2 (a). The number of causal states for $\epsilon - machine$ provides a rough reflection of the complexity of the user's behavior because each causal state is a "further refinement of the past for predictive sufficiency" (Darmon, 2015, p. 125). Since the majority of the users had only one state, indicating that their trolling behavior is like flipping a coin that whether troll or not is not a decision made based on their own past behavior(s). For most users, self-driven trolling behavior is not a complex process. Figure 4.2(b) presents the distribution of the maximum history lengths used for the estimated models, showing 455 (55.13%) users had the maximum history length of one, 212 (25.64%) users had the maximum length of four, and 67 (8.12%) users had the maximum length of five. The maximum history length suggests how many past behaviors have been used to predict future behaviors. Thus, based on the maximum lengths of the $\epsilon - machines$, more than half of the users only remembered the instant past behavior.

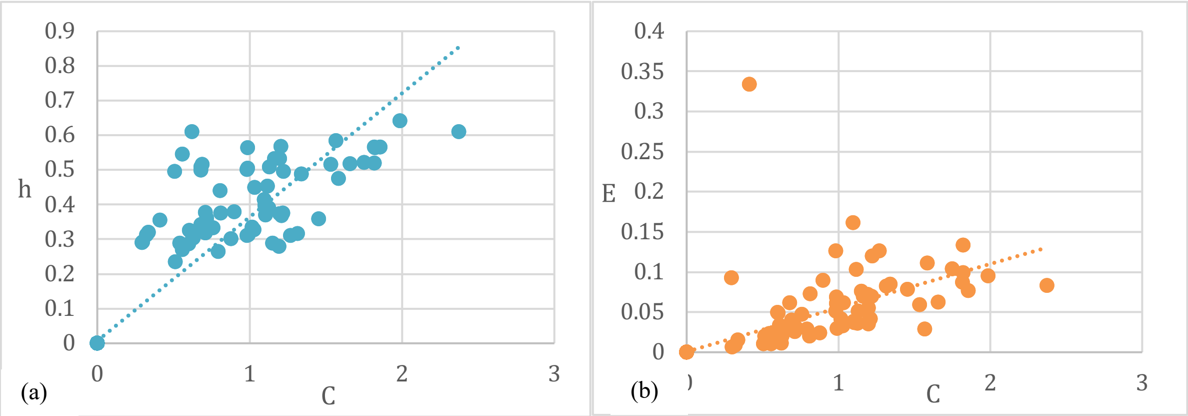
Figure 4.2. *Distribution of States and Maximum History Length in the Self-Driven Models.*



Note. (a) The distribution of states of $\epsilon - machines$. (b) The distribution of maximum history length of $\epsilon - machines$.

Secondly, we investigated the complementary information theoretic measures of the predictive state models. Figure 4.3 (a) and (b) shows the relations between predictive complexity C, remaining uncertainty h, and predictable information E. Unsurprisingly, both remaining uncertainty h (*r(826)*=0.942, *p*<0.001) and predictable information E (*r(826)*=0.905, *p*<0.001) are highly correlated with predictive complexity C. Models which require more stored information for prediction are more likely to have high rate of remaining uncertainty and high predictable information. In short, more complex processes communicate more from the past to the future, but at the same time, the probability of prediction error is also higher.

Figure 4.3. *Predictive Complexity vs. Remaining Uncertainty and Predictable Information with Regression lines of Self-Driven Models*



Note. (a) Predictive complexity vs. remaining uncertainty (b) Predictive complexity vs. predictable information

Moreover, ANOVA tests were run to further explain the architecture of self-driven models between the information theoretic measures and the number of states and the maximum history length L_{max}. Specifically, ANOVA tests were run on models with more than one states, since a one-state self-driven model basically follows binomial distribution that the past and the future are independent and share no information. Presented in Table 4.1., the results showed that models with more states are more likely to have larger predictive complexity and remaining uncertainty, and the models with longer history length are more likely to have higher remaining uncertainty. It echoed the investigation of information theoretic measures that a complex self-driven model with long memory is more unpredictable.

Table 4.1. *ANOVA results for the Predictable Information, Predictive Complexity, and Remaining Uncertainty of Self-Driven Models (N=113)*

	Effect	*F*	
C	Number of States	17.973	***
	L_{max}	3.539	
E	Number of States	0.003	

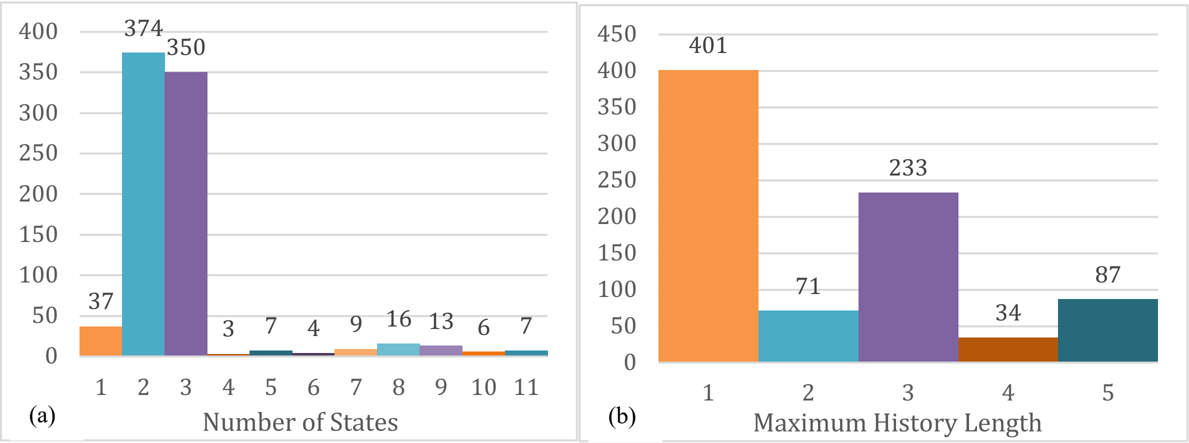
	L_{max}	1.878	
h	Number of States	20.198	***
	L_{max}	8.405	**

Note. ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

4.3.2. Social-driven Models

$\epsilon - transducer$ models were then estimated with parent time-series inputs, resulting in a majority of models with two causal states (374 users, 45.32%) or three causal states (350 users, 42.43%). The largest number of states detected from the models was eleven (7 users, 0.85%). The distribution is shown in Figure 4.4(a). Most of the users’ trolling behaviors can be described by a 2-state or 3-state $\epsilon - transducers,$ suggesting there is a hidden pattern guiding most users troll or not troll. Furthermore, for some other users with more states, the hidden patterns are more complex and difficult to understand. In addition, Figure 4.4(b) shows the distribution of maximum history length for $\epsilon - transducers.$ Similar to the one for self-driven models, around half of the users (401 users, 48.54%) had a maximum history length of one, followed by the maximum length of three with 233 users (28.21%) and the maximum length of five with 87 users (10.53%). Again, half of the users only remembered the instant past interaction with others, and the other half might have a better memory that can date back more interactions.

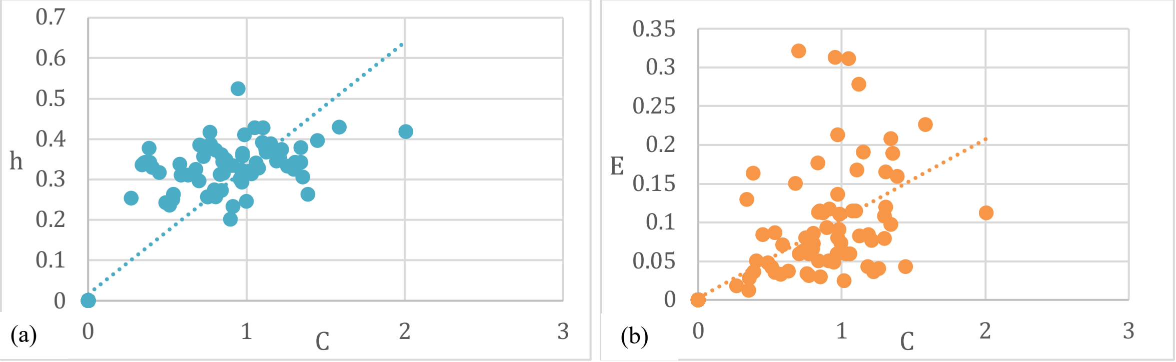
Figure 4.4. *Distribution of States and Maximum History Length in the Social-Driven Models*



Note. (a) The distribution of states of $\epsilon - machines$. (b) The distribution of maximum history length of $\epsilon - transducer$.

The complementary information theoretic measures of the predictive state models then were investigated. Figure 4.5 (a) and (b) shows the relations between predictive complexity C, remaining uncertainty h, and predictable information E. Remaining uncertainty h (*r(826)*=0.922, *p*<0.001) and predictable information E (*r(826)*=0.769, *p*<0.001) are both significantly correlated with predictive complexity C. Similar to Self-driven models, more complex social-driven models are more likely to have high rate of remaining uncertainty and more predictable information communicated from the past to the future.

Figure 4.5. *Predictive Complexity vs. Remaining Uncertainty and Predictable Information with Regression lines of Social-Driven Models*



Note. (a) Predictive complexity vs. remaining uncertainty (b) Predictive complexity vs. predictable information

To further understand the structure of social-driven models between the information theoretic measures and the number of states and the maximum history length L_{max}, ANOVA tests were employed. Presented in Table 4.2., the results showed that the number of states and maximum history length of the models are significantly associated with the information theoretic measures. Models with more states and longer history length are more likely to have larger predictive complexity, predictable information, and remaining uncertainty, which indicates that a complex model with long memory provides more information for prediction but at the same time is more unpredictable.

Table 4.2. *ANOVA results for the Predictable Information, Predictive Complexity, and Remaining Uncertainty of Social-Driven Models*

	Effect	F	
C	Number of States	61.016	***
	L_{max}	19.442	***
E	Number of States	21.853	***
	L_{max}	5.265	*
h	Number of States	79.270	***
	L_{max}	28.973	***

Note. ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

4.3.3. Self-driven vs Social-driven Models

To address RQ 4.1, we applied a Welch's t-test showing a significant difference in the number of states between the social-driven and self-driven models (t = 7.947, df = 824, p < 0.0001). The social-driven models had significantly more states than the self-driven models. In addition, a Welch's t-test was applied on predictable information E between self-driven and social-driven models (t = -5.615, df = 824, p < 0.0001), indicating that social-driven models provide significantly more information for predicting the future. The social-driven models provide additional information about users’ trolling behaviors compared with the self-driven models in general. Although it seems that trolling behaviors are haphazardly based on their own previous behaviors for most users, the social-driven models reveal a complex hidden pattern guiding the behavior. Both self and social inputs provide information to predict the future behaviors, but the social-driven model describe more about the trolling process. Trolling behavior is self-social-together-motivated.

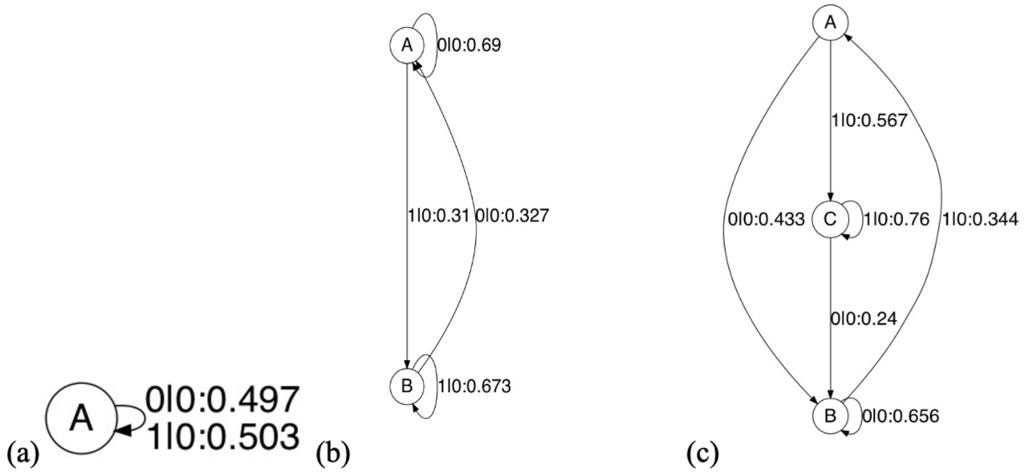
RQ 4.2 asked whether trolling behavior is contagious or not. Testing contagion is testing whether a trolling parent comment will lead to a trolling comment. As in the social-driven models, one state suggests that trolling behavior is more likely to be a random event, while multiple states indicate hidden patterns behind the behaviors. A Welch's t-test was applied for the number of trolling messages between users with multiple-state social-driven model and one-state social-driven model, and the result (t = 0.717, df = 824, p < 0.05) shows that users with hidden patterns are more likely to troll, supporting that users are socially influenced by the previous trolling messages from others.

To answer RQ 4.3, multiple predictive states in the self-driven and social-driven models unfold the hidden patterns of trolling behaviors. There are 109 (13.25%) users with a multiple-state self-driven model and 790 (95.57%) users with a multiple-state social-driven model in our dataset. Hidden patterns can be found in most co-dynamic systems of trolling behaviors, though with the increasing number of states, it can be very difficult to interpret the states and the dynamics. Moreover, a Welch's t-test was applied for the predictive complexity between self-driven models and social-driven models, and the result (t = =3.883, df = 824, p < 0.001) supporting that trolling behavior is a social influenced complex dynamic with sophisticated hidden structures. In addition, an exploratory test of the number of comments between one-state social-driven models and multiple-state social-driven models showed that activeness was not related to the dynamic patterns of trolling behaviors (t = 0.381, df = 824, p = 0.704).

4.3.4. Case Studies

To better illustrate the dynamic pattern of the self-driven and social-driven models, we provide case studies for models with different states. Figure 4.6 (a) shows the $\epsilon - machine$ for one of the one-state users. For users with only one state, troll or not troll is like flipping a coin that it is not a decision made based on their own past behavior. Whether they acted as a troll is out of random or, at least, for the reason that cannot be detected from the self-driven model. There is no obvious pattern detected from their own behaviors. Whereas in Figure 4.6 (b) is the $\epsilon - machine$ for one of the two-state users. Their behavioral pattern distinguishes between two states that we call a troll state (A) and a non-troll state (B), and they process memory for both stages: as they switch from non-troll to troll, they are more likely to remain troll, and *vice versa*. For this case, it is interesting to note that it is quite as likely to switch from trolling to non-trolling (31% and 32.7%). In addition, the three-state users' $\epsilon - machine$ is illustrated in Figure 4.5 (c). There is a stage in addition to the troll (C) and non-troll (B) state, and we call it the transition state (A). With a 56.7 % chance, this user will advance to trolling from the transition state, and a 43.3% chance to the non-troll state. Once the user arrives at the troll state, they will stay there with a probability of 76%. Note that the transition state (A) cannot be reached from the troll state (C) directly. It seems that the transition state serves the purpose of a threshold for becoming a troll or not. Being in a non-trolling state, the probability of staying there is quite high, 65.6 %. Being in the trolling state, the probability of staying a troll is even higher, 76%. Being in the transition state, thresholds are more evenly distributed, and the user seems to be in an uncertain and unstable situation of becoming a troll or not.

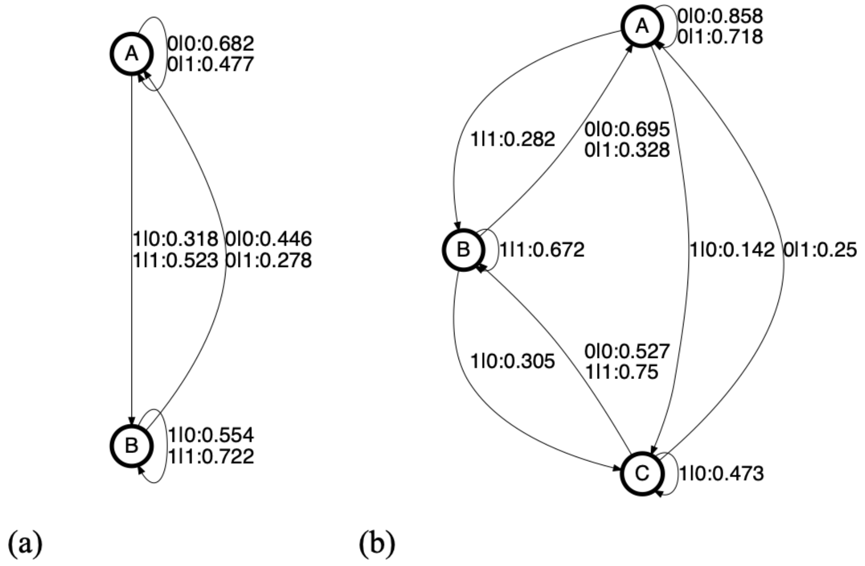
Figure 4.6. *The* $\epsilon - machines$ *for the users' trolling behavior*



Notes: Note, the input is always [$\dots$|0] (no social influence). (a) The one-state $\epsilon -$ machines. (b) The two-state $\epsilon -$ machines. (c) The three-state $\epsilon -$ machines.

Moving to the predictive state model that incorporates social inputs, the knowledge of the recent past of both their own and their social input (parent comment) behaviors provides sufficient information for predicting their future behavior. Figure 4.7(a) highlights a two-state user's epsilon-transducer. When the user's own previous behavior is a trolling behavior, the user is likely to switch to a troll state, and if the user is replying to a troll parent comment, it will reinforce the switch to the troll state and the user will stay in that state. Figure 4.7(b) is the epsilon-transducer for a user with three states. The user exhibits both self and social memory in the sense captured by the model. There are two routes from the non-troll state (A) to the troll state (B), one is directly from the non-troll state to the troll state, and the other is through the transition state (C). Unlike the transition state in the three-state epsilon-machine, the user can switch from the transition state to both troll and non-troll behavior.

Figure 4.7. *The* $\epsilon - transducers$ *for the users' trolling behavior*



Notes: (a) The two-state $\epsilon - \mathrm{transducer}$. (b) The three-state $\epsilon - \mathrm{transducer}$.

4.4. Discussion

We employed a unique, minimally complex, maximally predictive model to decompose the dynamic of trolling behaviors. As one of the most concerning online deviant behaviors, trolling is showed to irritate internet users, disrupt online discussions, and harm online communities. By approaching from a behavioral scientific perspective and dynamic system theory, we examined the hidden patterns of the behavioral models. The hidden pattern shows that trolling behavior is complex process that can be self-driven and social-driven, indicating internal influence and external influence for people engaging in trolling behaviors. In addition, compared to the self-driven models, more social-driven models provide more information for predicting the future behaviors and exhibit dynamical patterns, suggesting the contagion effect of the behavior.

One important finding of this study is that seeing trolling behaviors will encourage individuals to engage in trolling behaviors and reinforce the engagement. Consistent with previous research (Cheng et al., 2017; Shen et al., 2020), the complex interpersonal dynamics of trolling behavior revealed that trolling is contingent on the ongoing and simultaneous flow of communicative interactions with others. One possible explanation is the online disinhibition effect with the anonymous environment of Reddit and negative emotional contagion. Unlike other popular social media platforms such as Facebook or Twitter, Reddit encourages individuals not to use real identities (Reddit, 2021). Anonymity serves the purpose of physical, virtual, and emotional distancing between individuals, leading to toxic disinhibition. And the negative emotional contagion fuels the process of disinhibition. A negative mood can be transmitted across different discussions, leading people to emotional outbursts. However, such persistence was not found in individual dynamics of trolling behaviors. Though the individuals with multiple-state self-driven models in our case study showed a high probability of staying in the troll state, most users did not exhibit multiple probability states with persisting patterns. Unlike in social-driven models that individuals can be provoked by the parent comments, individuals are less likely to be internally motivated to engage trolling behaviors.

Furthermore, we unexpectedly found that activeness was not related to trolling behaviors. In contrast to growing concerns on the normalization of online deviant behaviors, we did not find active users who spent more time on Internet trolls more. Unlike the gaming environment (Shen et al., 2020), which requires an immediate response, online discussions may offer an opportunity for individuals to revisit their messages, deliberate their thoughts, and process others' messages. This thinking process can calm down negative emotions in the short term and can help develop critical thinking ability and accumulate knowledge in the long run. Actively engaging in an online discussion, individuals are likely to become more informed decision-makers regarding trolling behaviors. The finding may kindle the light of hope in a civil online deliberative discussion.

A practical implication that can be derived from the study is the importance of individuals. Preventing deviant behaviors from contaminating online communities is not solely responsible for platform designers, policymakers, or community moderators. It is widely recognized that platform policies and community norms on the aggregate level can tame trolls. At the same time, the study hints the critical role of individuals, who can easily be influenced and carry out trolling behaviors. In addition to strategies on the aggregate level, individuals level strategies are also an essential to intervene in the development of trolling and other online deviant behaviors. Even though trolling behavior is complex process and hard to predict, different reaction strategies can be taken to jump out of "trolling loop". For instance, individuals can implement the idea of "Don't feed the troll". Instead of responding with emotional outbursts immediately, individuals can first critically analyze and reflect the messages to see if it is worth replying. Therefore, the effort aiming at individual level intervention should also go into combating online deviant behaviors.

4.4.1. Limitation and Future Work

Few limitations should be kept in mind about the study when we interpret the findings. First, our findings may not be generalizable for other social media platforms, such as YouTube or Twitter. As Reddit has its own culture, especially, they take a different approach to controversial messages and communities. Their anti-harassment policy relies heavily on sub-reddit moderators, leading to a different treatment of the same message in different subreddits. This is also related to our second limitation on the definition of trolling. Trolling, or broadly, online deviant behaviors, is a complex concept that people may perceive differently contextualwise, leading to various reactions reflected in our dataset. Third, one important assumption is that our time series is a discrete-time stochastic process, while the duration between two time points may vary. The variation may further complicate the mathematical structures of trolling behaviors. Further studies may also take seasonality into consideration. In addition, further studies can also focus on investigate how different emotion displayed in the comments affect trolling behavior and stochastic models of the behavior.

Despite the limitations, our research provides a new, behavioral scientific approach to communication behaviors, where communication is considered a dynamic system with hidden structures. From such a perspective, we revealed the hidden structures of trolling behaviors and uncovered the social motivation of the behavior. Future work can address the different behavioral dynamics, digging into the external force within various contexts and environments of the overall online sphere. With social media becoming the major source and platform for news and democratic deliberation, understanding the complex dynamics of trolling and other online deviant behaviors is crucial for developing a safe and healthy discussion place.

Chapter 5. Conclusion

Social media were credited for the potential to revive trust in public space by providing opportunities for social and political participation, but the hope has been shadowed by the concerning phenomenon of online deviant behaviors. As shown in recent studies (Antoci et al., 2019; Shmargad et al., 2021), online deviant behaviors such as incivility and trolling are a dynamic and normative process that people are gradually seeing as the norm of online interactions. As the phenomenon runs rampant in online space, scholars and observers are concerned about its contribution to the erosion of democracy (Anderson & Rainie, 2020). The behaviors have been shown to deteriorate the discussion quality (Rainie et al., 2017), decrease social engagement (Kim & Park, 2019), salient minority perspectives (Ordoñez & Nekmat, 2019), affect individuals' emotions and cognitions (Chen & Ng, 2017), and further influence their behaviors (Cheng et al., 2017).

Even though extensive research has been dedicated to the field, it is not clear about the dynamic of the behavior on different levels. First, despite the general observation of the ubiquity of online deviant behaviors, the dynamics within and between contexts are largely unknown. Whether and how discussion topics and external features intersect with deviant behaviors remain to be tested. Second, little do we know the role of online communities in the dynamic of deviant behaviors. The nature of the community network may hatch perceptions and normative actions towards deviant behaviors, and how such community practices get involved in the mechanism of deviant behaviors requires scholarly attention. Third, the previous research showed that ordinary people can be triggered into engaging in deviant behaviors by the discussion contents and emotions (Cheng et al., 2017), but little has been known about the mechanism of behavior adaptation during the communication process.

5.1. Summaries of the Dissertation Studies

This dissertation research makes an important contribution to understanding the dynamics of online deviant behaviors on social media platforms from the macro, meso, and micro levels. It offers a thorough discussion from contextual, network, and behavioral perspectives to clarify the process of how deviant behaviors are embedded in individual processing, community norms, and platform contexts. It provides empirical evidence on how deviant behaviors developed over time in online space, how communities tickle the problematic behaviors, and how individuals adapt the behaviors.

First, our incivility data from Reddit suggested that, unlike the general observation, even though the volume of incivility increased during the past decade on the social media platform, the proportion stayed around 10% of all discussions. The variations were found among different topics and groups, and the fluctuations were found according to external events. The political discussion generally entailed more uncivil comments than groups whose purpose is not political-related but naturally involves socio-political discussions and groups for non-political topics. But that does not apply to geek culture. Gaming groups that have discussions around both political and gaming issues, generated the highest proportion of incivility, due to the special culture that encourages deviant behaviors. In addition, the mix of political leanings does not derive more uncivil comments compared with uniform-leaning groups, while groups that are not designated for politics are more likely to facilitate civil discussion on politics than groups with a political purpose.

One practical implication from the study as we observed, is that platform policies do play a role in calming down incivility. As Reddit imposed the anti-harassment policy in 2015, a dramatic decrease in incivility had been witnessed, which also affected later years. While debates about whether social media platforms should take action on preventing deviant behaviors are still going on, Reddit empowered its users to identify and report harassing messages. As Chen et al. (2019) indicated, there is no universal definition of incivility nor clear-cut categories for deviant behaviors. An imposed definition from platforms will inevitably result in a particular worldview forced on the users. Thus, aggregated reporting systems can generate more opinions on deviant behaviors, which may help individual users reach a mutual understanding of such behaviors, further leading to healthy social norms in the online environment.

Second, by investigating pre-troll and post-troll commenting networks of YouTube videos, the study showed that the community's network structure and individuals' positions imply their closeness with the community, which has a determinant role in how individuals will treat deviant behaviors and those who engaged in such behaviors. In a well-connected community where individuals are close to each other, they are more likely to see trolling as an attack that will trigger their counterattacks, which means protecting the community. In addition, similar reactions may happen to individuals who posit in the center of the communicating network of a community and are actively engaged in interactions. Community norms also serve as a guide for individuals, where individuals follow the collective response to trolling behaviors in the community.

Practically, as we argued previously that online deviant behaviors can be considered as a normative process, establishing and reinforcing appropriated community norms is essential to a healthy online community. The study suggested that responses to trolls in an online community will only beget more responses, leading to meaningless discussions which waste time and resources. To avoid that, a non-response norm should be established. For instance, set up rules about what counts as deviant behaviors and how individuals in the communities should react to them, then everyone in the community has the responsibility to follow the rules and educate the newcomers. During such a process, the community can normalize the non-response treatment which may counter the normative process of deviant behaviors.

Third, the predictive models of active users on Reddit show that there is a complex hidden pattern guiding individuals' online behaviors. Engaging in deviant behaviors can be self-motivated, social-motivated, and self-social-together-motivated, and social-motivated and self-social-together-motivated models provide more information for how individuals engage in deviant behavior. The comments individuals replied to are important as those trigger a responding behavior, and likely to lure individuals to respond in deviant ways. In addition, counter-intuitively, if someone is more active in online space does not necessarily mean that they are more likely to be influenced by others' deviant behaviors. More likely, with the time spent and activities participated in online, active individuals accumulated knowledge and established critical thinking processes that immunize them from deviant behaviors.

Again, the study suggests that individuals can be influenced by the online environment and by the comments they have come across. Despite the practical interventions involved in platform-level policies, user reporting systems, or community-level regulations, individual level strategy is in need for preventing the spread of deviant behaviors. For example, by analyzing and evaluating posts, comments, and messages, individuals may be more likely to calm down and reflect on what they have read and what they will post. Also, individuals can develop an awareness of different ways of responding. Replying to a trolling message is not the only way of responding while utilizing reporting systems, outing trolls, and even ignoring trolls can also serve the purpose of combating trolling.

5.2. Limitations

Notwithstanding the contributions, several limitations need to be addressed and kept in mind for future research. The first major limitation across all three studies is the universal identification of online deviant behaviors. As shown in the studies, deviant behaviors are highly contextual. Though scholars want to offer conceptual clarity and an operational approach, such as differentiating incivility and impoliteness, developing machine classifiers based on universal definition, etc., individuals hold different views on how to understand incivility, trolling, and other deviant behaviors. Despite content itself, individual characteristics also matter in how individuals understand deviant behaviors. For instance, individuals who are older (Klempka & Stimson, 2014) and who score high on Agreeableness in Big Five Personality (Rains et al., 2017) are more likely to perceive a comment as uncivil. Individuals' social identity, like partisanship, also shapes their views on incivility. Therefore, a more sophisticated approach to identifying deviant behaviors is needed. How to composite contextual factors, individual characteristics, and message features to form a complex definition of online deviant behaviors for research purposes and practical purposes is a task for future research.

Second, all three studies used observational data and the lack of other data collection methods may limit or bias the findings. Although the first study used longitudinal data involving 11 years, no causal relationship was tested. Similarly, the second study, which used cross-sectional data, failed to make a causal inference. In addition, the large datasets in all three studies provide sufficient information about individuals' online behaviors, they may also raise the problem of multicollinearity, where the factors are correlated with each other. Another practical issue during data collection is that the deleted or removed comments were not accessible, thus the results may be biased. It is highly likely that the deleted or removed comments contain certain types of deviance, which may lead to underestimations of the deviant behaviors.

Third, generalizability is limited in the dissertation studies. As all three studies focused on one social media platform, namely Reddit and YouTube, the findings may not be generalizable to other social media platforms. Especially, Reddit and YouTube have their own target users, community structures, and policies for harassment. In addition, for the second and third studies, the data was collected within a specific period, which included the 2016 presidential election, leading to a possibility for a surge of deviant behaviors online. The focus of the two studies was also laid upon political discussions, this, the results may not generalize to other time periods and other domains. Future research can carry out cross-platform work, as well as works focusing on the different time periods and domains like gaming, health, and entertainment.

5.3. Future Directions

The dissertation research investigated online deviant behaviors from the macro, meso, and micro levels, and the studies' limitations discussed above pointed to some promising future research directions, including the definitions and identifications, and research methods. Besides those, the communication dynamic involving other online factors should also be paid scholarly attention.

Definitions and identifications, as mentioned multiple times in the dissertation cannot be universal. There is a great variation in what individuals believe is deviant, and categorization of deviant behaviors relies on the understanding of deviance in a particular culture. And the various definitions also lead to different reactions. Instead of defining deviant behaviors from the eyes of beholders, future research can embrace the diversity of perspectives and employ definitions from online communities. Furthermore, as the use of multimedia, such as memes and short videos, becomes more and more prevalent, more inclusive definitions or sub definitions should be proposed to adjust to the fast-paced computer-mediated environment. Current definitions of different types of deviant behaviors are majorly textualized, focusing on text and content of messages and comments, while other formats of the behaviors should also be considered. In addition, behavioral characteristics, such as posting frequencies and the length of messages, should be taken into account when defining deviant behavior in future research.

Method-wise, the dissertation relies heavily on observational data, while surveys and experiments can serve the purposes to fill out the gaps in understanding individual members in an online community, establishing causal relationships, and explaining the missing part of the picture. For example, to define deviant behaviors, surveys can help with having individual members in an online community report what is deviant in their particular conversations. On the other hand, experiments are able to identify antecedents and consequences. Moreover, the triangulation method can be carried out to different platforms, time periods, and domains in future research.

While the focus of this dissertation is online deviant behaviors, there is more to investigate how these types of behaviors are embedded in the CMC environment. Questions could be asked if there are any other possible antecedents that will lead to deviant behaviors. Developing from the second study, the network perspective could offer a look into whether and how different network structures and positions may lead to deviant behaviors. Research could take an approach on the level of community and organization to test how organizational features and structural characteristics could be antecedents of deviant behaviors. The organizational perspective could also be applied to research the effects and consequences of deviant behaviors.

Current research on the effects is still mainly on the level of individuals and little research has been done to understand the normalization process of online deviant behaviors. In addition, online deviant behaviors are said to be attributed to artificial intelligence (AI) as a powerful emerging technology. Future studies can emphasize different forms of AI, such as search algorithms, recommendation algorithms, and social bots, and how AI promotes or moderates online deviant behaviors. Furthermore, online deviant behaviors are connected with other important topics in modern civil deliberation, such as polarization, and misinformation. How dynamically those topics interplay together affecting human attitudes and behaviors and affecting society at large is also an important subfield that needs scholarly attention.

REFERENCE

Ackland, R., & O'neil, M. (2011). Online collective identity: The case of the environmental movement. *Social Networks, 33*(3), 177-190.

Akers, R. L., Krohn, M. D., Lanza-Kaduce, L., & Radosevich, M. (1995). Social learning and deviant behavior: A specific test of a general theory. In *Contemporary Masters in Criminology* (pp. 187-214). Springer.

Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior, 63*, 433-443.

Alexa. (2019). *Top Sites in United States.* Retrieved December 14 from http://www.alexa.com/topsites/countries/US

Alonzo, M., & Aiken, M. (2004). Flaming in electronic communication. *Decision Support Systems, 36*(3), 205-213.

Alsmadi, I., & O'Brien, M. J. (2020). How Many Bots in Russian Troll Tweets? *Information Processing & Management, 57*(6), 102303.

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "nasty effect:" Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication, 19*(3), 373-387.

Anderson, A. A., & Huntington, H. E. (2017). Social media, science, and attack discourse: How Twitter discussions of climate change use sarcasm and incivility. *Science Communication, 39*(5), 598-620.

Anderson, A. A., Yeo, S. K., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2018). Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research, 30*(1), 156-168.

Anderson, J., & Rainie, L. (2020). Many tech experts say digital disruption will hurt democracy. *Pew Research Center. Internet & Technology. Feb, 21*.

Ansong, E. D., Takyi, T., Damoah, D., Ampomah, E. A., & Larkotey, W. (2013). Internet trolling in Ghana. *International Journal of Emerging Science and Engineering, 2*(1), 42-43.

Antoci, A., Bonelli, L., Paglieri, F., Reggiani, T., & Sabatini, F. (2019). Civility and trust in social media. *Journal of Economic Behavior & Organization, 160*, 83-99.

Baer, D. (2021). *The 31 Biggest Subreddits*. Retrieved June 25 from https://blog.oneupapp.io/biggest-subreddits/

Baker, P. (2001). Moral panic and alternative identity construction in Usenet. *Journal of Computer-Mediated Communication, 7*(1), JCMC711.

Ballard, M. E., & Welch, K. M. (2017). Virtual warfare: Cyberbullying and cyber-victimization in MMOG play. *Games and culture, 12*(5), 466-491.

Bandura, A., & Walters, R. H. (1977). *Social learning theory* (Vol. 1). Prentice-hall Englewood Cliffs, NJ.

Barberá, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., & Tucker, J. A. (2019). Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review, 113*(4), 883-901.

Barnett, N., & Crutchfield, J. P. (2015). Computational Mechanics of Input–Output Processes: Structured Transformations and the $$\epsilon $$ ϵ-Transducer. *Journal of statistical physics, 161*(2), 404-451.

Basmmi, A. B. M. N., Abd Halim, S., & Saadon, N. A. (2020). Comparison of Web Services for Sentiment Analysis in Social Networking Sites. IOP Conference Series: Materials Science and Engineering,

Bauman, S., Toomey, R. B., & Walker, J. L. (2013). Associations among bullying, cyberbullying, and suicide in high school students. *Journal of adolescence, 36*(2), 341-350.

Beres, N. A., Frommel, J., Reid, E., Mandryk, R. L., & Klarkowski, M. (2021). Don’t you know that you’re toxic: Normalization of toxicity in online gaming. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems,

Bergami, M., & Bagozzi, R. P. (2000). Self-categorization, affective commitment and group self-esteem as distinct aspects of social identity in the organization. *British journal of social psychology, 39*(4), 555-577.

Bergstrom, K. (2011). “Don’t feed the troll”: Shutting down debate about community expectations on Reddit. com. *First Monday, 16*(8).

Berry, J. M., & Sobieraj, S. (2013). *The outrage industry: Political opinion media and the new incivility*. Oxford University Press.

Bialek, W., Nemenman, I., & Tishby, N. (2001). Predictability, complexity, and learning. *Neural computation, 13*(11), 2409-2463.

Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction, 1*(CSCW), 1-19.

Borah, P. (2013). Interactions of news frames and incivility in the political blogosphere: Examining perceptual outcomes. *Political Communication, 30*(3), 456-473.

Breeze, M. (2012). *The problems with anonymous trolls and accountability in the digital age.* https://thenextweb.com/insider/2012/10/27/the-problems-with-anonymous-trolls-and-accountability-in-the-digital-age/

Brown, J., Broderick, A. J., & Lee, N. (2007). Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of interactive marketing, 21*(3), 2-20.

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and individual Differences, 67*, 97-102.

Buntain, C., Bonneau, R., Nagler, J., & Tucker, J. A. (2021). YouTube recommendations and effects on sharing across online social platforms. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1-26.

Cameron, J. E. (2004). A three-factor model of social identity. *Self and identity, 3*(3), 239-262.

Cappella, J. N. (1979). Talk-silence sequences in informal conversations I. *Human Communication Research, 6*(1), 3-17.

Centola, D. (2010). The spread of behavior in an online social network experiment. *science, 329*(5996), 1194-1197.

Chen, G. M., & Ng, Y. M. M. (2017). Nasty online comments anger you more than me, but nice ones make me as happy as you. *Computers in Human Behavior, 71*, 181-188.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing,

Christensen, P. N., Rothgerber, H., Wood, W., & Matz, D. C. (2004). Social norms and identity relevance: A motivational approach to normative behavior. *Personality and Social Psychology Bulletin, 30*(10), 1295-1309.

Christopherson, K. M. (2007). The positive and negative implications of anonymity in Internet social interactions:“On the Internet, Nobody Knows You’re a Dog”. *Computers in Human Behavior, 23*(6), 3038-3056.

Chun, J., Lee, J., Kim, J., & Lee, S. (2020). An international systematic review of cyberbullying measurements. *Computers in Human Behavior*, 106485.

Cicchirillo, V., Hmielowski, J., & Hutchens, M. (2015). The mainstreaming of verbally aggressive online political behaviors. *Cyberpsychology, Behavior, and Social Networking, 18*(5), 253-259.

Civility in America 2019: Solutions For Tomorrow. (2019). Weber Shandwick. https://www.webershandwick.com/wp-content/uploads/2019/06/CivilityInAmerica2019SolutionsforTomorrow.pdf

Clarke, I. (2019). Functional linguistic variation in Twitter trolling. *International Journal of Speech Language and the Law, 26*(1), 57-84.

Clement, J. (2019). *Share of adult internet users in the United States who have personally experienced online harassment as of December 2018*. https://www.statista.com/statistics/333942/us-internet-online-harassment-severity/

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication, 64*(4), 658-679.

Coles, B. A., & West, M. (2016). Trolling the trolls: Online forum users constructions of the nature and properties of trolling. *Computers in Human Behavior, 60*, 233-244.

Cook, C. L. (2019). Between a troll and a hard place: the demand framework's answer to one of gaming's biggest problems. *Media and Communication, 7*(4), 176-185.

Cortina, L. M., Magley, V. J., Williams, J. H., & Langhout, R. D. (2001). Incivility in the workplace: incidence and impact. *Journal of occupational health psychology, 6*(1), 64.

Cover, R. (2012). Performing and undoing identity online: Social networking, identity theories and the incompatibility of online profiles and friendship regimes. *Convergence, 18*(2), 177-193.

Croes, E. A., Antheunis, M. L., Schouten, A. P., & Krahmer, E. J. (2016). Teasing apart the effect of visibility and physical co-presence to examine the effect of CMC on interpersonal attraction. *Computers in Human Behavior, 55*, 468-476.

Crosslin, K., & Golman, M. (2014). "Maybe you don't want to face it"–College students' perspectives on cyberbullying. *Computers in Human Behavior, 41*, 14-20.

Crutchfield, J. P. (1994). The calculi of emergence: computation, dynamics and induction. *Physica D: Nonlinear Phenomena, 75*(1-3), 11-54.

Crutchfield, J. P. (2017). The origins of computational mechanics: A brief intellectual history and several clarifications. *arXiv preprint arXiv:1710.06832*.

Crutchfield, J. P., Ellison, C. J., & Mahoney, J. R. (2009). Time's barbed arrow: Irreversibility, crypticity, and stored information. *Physical review letters, 103*(9), 094101.

Crutchfield, J. P., & Feldman, D. P. (1997). Statistical complexity of simple one-dimensional spin systems. *Physical Review E, 55*(2), R1239.

Crutchfield, J. P., & Young, K. (1989). Inferring statistical complexity. *Physical review letters, 63*(2), 105.

Cruz, A. G. B., Seo, Y., & Rex, M. (2018). Trolling in online communities: A practice-based theoretical perspective. *The information society, 34*(1), 15-26.

Curry, D. (2021). *Reddit Revenue and Usage Statistics.* Retrieved July 12 from https://www.businessofapps.com/data/reddit-statistics/

Darmon, D. (2015). *Statistical methods for analyzing time series data drawn from complex social systems* University of Maryland, College Park].

Davidson, S., Sun, Q., & Wojcieszak, M. (2020). Developing a New Classifier for Automated Identification of Incivility in Social Media. Proceedings of the Fourth Workshop on Online Abuse and Harms,

de Mesquita Neto, J. A., & Becker, K. (2018). Relating conversational topics and toxic behavior effects in a MOBA game. *Entertainment computing, 26*, 10-29.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Donath, J. S. (2002). Identity and deception in the virtual community. In *Communities in cyberspace* (pp. 37-68). Routledge.

Druckman, J. N., Gubitz, S., Lloyd, A. M., & Levendusky, M. S. (2019). How incivility on partisan media (de) polarizes the electorate. *The Journal of Politics, 81*(1), 291-295.

Eady, G., Nagler, J., Guess, A., Zilinsky, J., & Tucker, J. A. (2019). How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *Sage Open, 9*(1), 2158244019832705.

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *science, 221*(4616), 1208-1210.

Ellis, D. G., & Fisher, B. A. (1975). Phases of conflict in small group development: A Markov analysis. *Human Communication Research, 1*(3), 195-212.

Febriana, S. (2019). Cyber Incivility Perpetrator: The Influenced of Dissociative Anonimity, Invisibility, Asychronicity, and Dissociative Imagination. Journal of Physics: Conference Series,

Feldman, L., Stroud, N. J., Bimber, B., & Wojcieszak, M. (2013). Assessing selective exposure in experiments: The implications of different methodological choices. *Communication Methods and Measures, 7*(3-4), 172-194.

Ferree, M. M., Gamson, W. A., Gerhards, J., & Rucht, D. (2002). Four models of the public sphere in modern democracies. *Theory and society, 31*(3), 289-324.

Fichman, P., & Sanfilippo, M. R. (2016). *Online trolling and its perpetrators: Under the cyberbridge*. Rowman & Littlefield.

Fisher, B. A., & Drecksel, G. L. (1983). A cyclical model of developing relationships: A study of relational control interaction. *Communications Monographs, 50*(1), 66-78.

Fletcher, R., & Nielsen, R. K. (2018). Are people incidentally exposed to news on social media? A comparative analysis. *New media & society, 20*(7), 2450-2468.

Fogel, A. (2006). Dynamic systems research on interindividual communication: The transformation of meaning-making. *Journal of developmental processes, 1*(1), 7-30.

Gervais, B. T. (2015). Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics, 12*(2), 167-185.

Gervais, B. T. (2017). More than mimicry? The role of anger in uncivil reactions to elite political incivility. *International Journal of Public Opinion Research, 29*(3), 384-405.

Graf, J., Erba, J., & Harn, R.-W. (2017). The role of civility and anonymity on perceptions of online comments. *Mass Communication and Society, 20*(4), 526-549.

Granovetter, M. (1973). The Strength of Weak Ties.-American Journal of Sociology. Vol. 78, Is. 6. P. 1360-1380.

Grassberger, P. (1986). Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics, 25*(9), 907-938.

Greenhow, C., & Robelia, B. (2009). Informal learning and identity formation in online social networks. *Learning, media and technology, 34*(2), 119-140.

Groshek, J., & Cutino, C. (2016). Meaner on mobile: Incivility and impoliteness in communicating contentious politics on sociotechnical networks. *Social Media+ Society, 2*(4), 2056305116677137.

Guess, A. M. (2021). (Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets. *American Journal of Political Science, 65*(4), 1007-1022.

Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior, 29*(3), 1159-1168.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current directions in psychological science, 2*(3), 96-100.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence),

Herrando, C., & Constantinides, E. (2021). Emotional contagion: a brief overview and future directions. *Frontiers in psychology, 12*, 712606.

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.

Hilvert-Bruce, Z., & Neill, J. T. (2020). I'm just trolling: The role of normative beliefs in aggressive behaviour in online gaming. *Computers in Human Behavior, 102*, 303-311.

Hinduja, S., & Patchin, J. W. (2007). Offline consequences of online victimization: School violence and delinquency. *Journal of school violence, 6*(3), 89-112.

Hogg, M. A. (2016). Social identity theory. In *Understanding peace and conflict through social identity theory* (pp. 3-17). Springer.

Hogg, M. A., & Reid, S. A. (2006). Social identity, self-categorization, and the communication of group norms. *Communication theory, 16*(1), 7-30.

Hong, F.-Y., & Cheng, K.-T. (2018). Correlation between university students' online trolling behavior and online trolling victimization forms, current conditions, and personality traits. *Telematics and Informatics, 35*(2), 397-405.

Howard, J. W. (2019). Free speech and hate speech. *Annual Review of Political Science, 22*, 93-109.

Huesmann, L. R., & Eron, L. D. (1984). Cognitive processes and the persistence of aggressive behavior. *Aggressive behavior, 10*(3), 243-251.

Hwang, H., Kim, Y., & Huh, C. U. (2014). Seeing is believing: Effects of uncivil online debate on political polarization and expectations of deliberation. *Journal of Broadcasting & Electronic Media, 58*(4), 621-633.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science, 22*, 129-146.

Kaur, A., & Chopra, D. (2016). Comparison of text mining tools. 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO),

Kelly, J. W., Fisher, D., & Smith, M. (2006). Friends, foes, and fringe: norms and structure in political discussion networks. Proceedings of the 2006 international conference on Digital government research,

Kenski, K., Coe, K., & Rains, S. A. (2020). Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research, 47*(6), 795-814.

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication, 71*(6), 922-946.

Kim, J. w., & Park, S. (2019). How perceptions of incivility and social endorsement in online comments (Dis) encourage engagements. *Behaviour & Information Technology, 38*(3), 217-229.

Kim, Y., & Gonzales, A. L. (2022). When we tolerate online incivility: Dual-Process Effects of Argument Strength and Heuristic Cues in Uncivil User Comments. *Computers in Human Behavior, 131*, 107235.

Klar, S. (2018). Neither Liberal nor Conservative: Ideological Innocence in the American Public. By Donald R. Kinder and Nathan P. Kalmoe. Chicago: University of Chicago Press, 2017. 224p. 26.00 paper. *Perspectives on Politics, 16*(3), 847-848.

Klar, S., & Krupnikov, Y. (2016). *Independent politics*. Cambridge University Press.

Klempka, A., & Stimson, A. (2014). Anonymous communication on the internet and trolling. *Concordia Journal of Communication Research, 1*(1), 2.

Knobloch-Westerwick, S., Mothes, C., & Polavin, N. (2020). Confirmation Bias, Ingroup Bias, and Negativity Bias in Selective Exposure to Political Information. *Communication Research, 47*(1), 104-124. https://doi.org/10.1177/0093650217719596

Komaç, G., & Çağıltay, K. (2019). An overview of trolling behavior in online spaces and gaming context. 2019 1st International Informatics and Software Engineering Conference (UBMYK),

Kou, Y. (2020). Toxic behaviors in team-based competitive gaming: The case of league of legends. Proceedings of the annual symposium on computer-human interaction in play,

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(24), 8788-8790.

KRC Research. (2018). *Civility in America 2018: Civility at Work and in Our*

Public Squares. KRC Research. Retrieved Dec 14 from https://www.webershandwick.com/wp-content/uploads/2018/06/Civility-in-America-VII-FINAL.pdf

Ksiazek, T. B. (2015). Civil interactivity: How news organizations' commenting policies explain civility and hostility in user comments. *Journal of Broadcasting & Electronic Media, 59*(4), 556-573.

Kunst, A. (2019). *How often do you see internet trolling on the following types of media?* https://www.statista.com/statistics/379997/internet-trolling-digital-media/

Kwak, H., Blackburn, J., & Han, S. (2015). Exploring cyberbullying and other toxic behavior in team competition online games. Proceedings of the 33rd annual ACM conference on human factors in computing systems,

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior, 28*(2), 434-443.

Lea, M., O'Shea, T., Fung, P., & Spears, R. (1992). *'Flaming'in computer-mediated communication: Observations, explanations, implications*. Harvester Wheatsheaf.

Lea, M., & Spears, R. (1991). Computer-mediated communication, de-individuation and group decision-making. *International journal of man-machine studies, 34*(2), 283-301.

Liang, H., & Zhang, X. (2021). Partisan Bias of Perceived Incivility and its Political Consequences: Evidence from Survey Experiments in Hong Kong. *Journal of Communication, 71*(3), 357-379.

Lin, Y. (2021). *10 Reddit Statistics Every Marketer Should Know in 2021*. Retrieved July 12 from https://www.oberlo.com/blog/reddit-statistics

Massanari, A. (2017). # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New media & society, 19*(3), 329-346.

Masullo Chen, G., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media+ Society, 5*(3), 2056305119862641.

McCosker, A. (2014). Trolling as provocation: YouTube's agonistic publics. *Convergence, 20*(2), 201-217.

McKenna, K. Y., & Green, A. S. (2002). Virtual group dynamics. *Group dynamics: theory, research, and practice, 6*(1), 116.

McKenna, K. Y., & Seidman, G. (2005). Social identity and the self: Getting connected online. *Cognitive technology: Essays on the transformation of thought and society*, 89-110.

Merritt, E. (2012). *An analysis of the discourse of Internet trolling: A case study of Reddit. com*

Millen, D. R., & Patterson, J. F. (2003). Identity disclosure and the creation of social capital. CHI'03 extended abstracts on Human factors in computing systems,

MonkeyLearn. (2020). *Sentiment Analysis*. MonkeyLearn. https://monkeylearn.com/sentiment-analysis/

Moor, P. J., Heuvelman, A., & Verleur, R. (2010). Flaming on youtube. *Computers in Human Behavior, 26*(6), 1536-1546.

Muddiman, A. (2017). Personal and public levels of political incivility. *International Journal of Communication, 11*, 21.

Muddiman, A. (2021). Conservatives and incivility. *Conservative Political Communication How Right-Wing Media and Messaging (Re) Made American Politics. Routledge*.

Murphy, S. C. (2004). 'Live in Your World, Play in Ours': The Spaces of Video Game Identity. *Journal of Visual Culture, 3*(2), 223-238. https://doi.org/10.1177/1470412904044801

Mutz, D. C. (2015). *In-your-face politics*. Princeton University Press.

Mutz, D. C., & Reeves, B. (2005). The new videomalaise: Effects of televised incivility on political trust. *American Political Science Review, 99*(1), 1-15.

Nevin, A. D. (2015). Cyber-Psychopathy: Examining the relationship between dark E-personality and online misconduct.

Nicholson, J., & Higgins, G. E. (2017). Social structure social learning theory: Preventing crime and violence. In *Preventing crime and violence* (pp. 11-20). Springer.

Nithyanand, R., Schaffner, B., & Gill, P. (2017a). Measuring offensive speech in online political discourse. 7th USENIX workshop on free and open communications on the internet (FOCI 17), Nithyanand, R., Schaffner, B., & Gill, P. (2017b). Online political discourse in the Trump era. *arXiv preprint arXiv:1711.05303*.

Ochs, E. (1993). Constructing social identity: A language socialization perspective. *Research on language and social interaction, 26*(3), 287-306.

Olson, C. S. C., & LaPoe, V. (2017). "Feminazis,""libtards,""snowflakes," and "racists": Trolling and the Spiral of Silence effect in women, LGBTQIA communities, and disability populations before and after the 2016 election. *The Journal of Public Interest Communications, 1*(2), 116-116.

Ordoñez, M. A. M., & Nekmat, E. (2019). "Tipping point" in the SoS? Minority-supportive opinion climate proportion and perceived hostility in uncivil online discussion. *New media & society, 21*(11-12), 2483-2504.

Ortiz, S. M. (2020). Trolling as a collective form of harassment: an inductive study of how online users understand trolling. *Social Media+ Society, 6*(2), 2056305120928512.

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New media & society, 6*(2), 259-283.

Pearson, E. (2009). All the World Wide Web'sa stage: The performance of identity in online social networks. *First Monday, 14*(3).

Peteraf, M., & Shanley, M. (1997). Getting to know you: A theory of strategic group identity. *Strategic Management Journal, 18*(S1), 165-186.

Pew Research Center. (2021a). *Social Media Use in 2021*. Retrieved June 25 from https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Pew Research Center. (2021b). *The State of Online Harassment*. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Poole, M. S. (2007). Generalization in process theories of communication. *Communication Methods and Measures, 1*(3), 181-190.

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-effects of computer-mediated communication. *Communication Research, 25*(6), 689-715.

Postmes, T., Spears, R., & Lea, M. (1999). Social identity, normative content, and" deindividuation" in computer-mediated groups.

Prior, M. (2007). *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press.

Raine, A. (2002). Annotation: The role of prefrontal deficits, low autonomic arousal, and early health factors in the development of antisocial and aggressive behavior in children. *Journal of Child psychology and Psychiatry, 43*(4), 417-434.

Rainie, H., Anderson, J. Q., & Albright, J. (2017). *The future of free speech, trolls, anonymity and fake news online*. Pew Research Center Washington, DC.

Rains, S. A., Kenski, K., Coe, K., & Harwood, J. (2017). Incivility and political identity on the Internet: Intergroup factors as predictors of incivility in discussions of news online. *Journal of Computer-Mediated Communication, 22*(4), 163-178.

Rajadesingan, A., Budak, C., & Resnick, P. (2021). Political discussion is abundant in non-political subreddits (and less toxic). Proceedings of the International AAAI Conference on Web and Social Media,

Ransbotham, S., Fichman, R. G., Gopal, R., & Gupta, A. (2016). Special section introduction—ubiquitous IT and digital vulnerabilities. *Information Systems Research, 27*(4), 834-847.

Reader, B. (2012). Free press vs. free speech? The rhetoric of "civility" in regard to anonymous online comments. *Journalism & mass communication quarterly, 89*(3), 495-513.

Reddit. (2021). *Reddit Policy*. Retrieved September 2 from https://www.reddithelp.com/hc/en-us/categories/360003246511-Privacy-Security

Reicher, S., Haslam, S. A., & Rath, R. (2008). Making a Virtue of Evil: A Five-Step Social Identity Model of the Development of Collective Hate. *Social and Personality Psychology Compass, 2*(3), 1313-1344. https://doi.org/10.1111/j.1751-9004.2008.00113.x

Rösner, L., Winter, S., & Krämer, N. C. (2016). Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior, 58*, 461-470.

Rossini, P. (2020). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 0093650220921314.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and social psychology review, 5*(4), 296-320.

Sadeque, F., Rains, S., Shmargad, Y., Kenski, K., Coe, K., & Bethard, S. (2019). Incivility detection in online comments. Proceedings of the eighth joint conference on lexical and computational semantics (* SEM 2019),

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Santana, A. D. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism practice, 8*(1), 18-33.

Seregina, A., & Schouten, J. W. (2017). Resolving identity ambiguity through transcending fandom. *Consumption Markets & Culture, 20*(2), 107-130.

Shalizi, C. R., & Crutchfield, J. P. (2001). Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of statistical physics, 104*(3), 817-879.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell system technical journal, 30*(1), 50-64.

Shen, C., Sun, Q., Kim, T., Wolff, G., Ratan, R., & Williams, D. (2020). Viral vitriol: Predictors and contagion of online toxicity in World of Tanks. *Computers in Human Behavior, 108*, 106343.

Shen, K. N., Yu, A. Y., & Khalifa, M. (2010). Knowledge contribution in virtual communities: accounting for multiple dimensions of social presence through social identity. *Behaviour & Information Technology, 29*(4), 337-348.

Shin, J. (2008). Morality and Internet Behavior: A study of the Internet Troll and its relation with morality on the Internet. Society for information technology & teacher education international conference,

Shmargad, Y., Coe, K., Kenski, K., & Rains, S. A. (2021). Social norms and the dynamics of online incivility. *Social Science Computer Review*, 0894439320985527.

Shores, K. B., He, Y., Swanenburg, K. L., Kraut, R., & Riedl, J. (2014). The identification of deviance and its impact on retention in a multiplayer game. Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing,

Silver, A., & Andrey, J. (2019). Public attention to extreme weather as reflected by social media activity. *Journal of Contingencies and Crisis Management, 27*(4), 346-358.

Sipser, M. (1996). Introduction to the Theory of Computation. *ACM Sigact News, 27*(1), 27-29.

Smith, P. K., Mahdavi, J., Carvalho, M., & Tippett, N. (2006). An investigation into cyberbullying, its forms, awareness and impact, and the relationship between age and gender in cyberbullying. *Research Brief No. RBX03-06. London: DfES.*

Sobieraj, S., & Berry, J. M. (2011). From incivility to outrage: Political discourse in blogs, talk radio, and cable news. *Political Communication, 28*(1), 19-41.

Sohn, D. (2009). Disentangling the effects of social network density on electronic word-of-mouth (eWOM) intention. *Journal of Computer-Mediated Communication, 14*(2), 352-367.

Squicciarini, A., Rajtmajer, S., Liu, Y., & Griffin, C. (2015). Identification and characterization of cyberbullying dynamics in an online social network. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015,

Stevens, H., Acic, I., & Taylor, L. D. (2021). Uncivil reactions to sexual assault online: linguistic features of news reports predict discourse incivility. *Cyberpsychology, Behavior, and Social Networking, 24*(12), 815-821.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems, 29*(4), 217-248.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior, 7*(3), 321-326.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification? China national conference on Chinese computational linguistics,

Sun, Q. (2022). Why Do You Stop Playing? Effects of Toxicity in MMOs *Unpublished manuscript*.

Sun, Q., & Shen, C. (2021). Who would respond to A troll? A social network analysis of reactions to trolls in online communities. *Computers in Human Behavior, 121*, 106786.

Sun, Q., Wojcieszak, M., & Davidson, S. (2021). Over-Time Trends in Incivility on Social Media: Evidence From Political, Non-Political, and Mixed Sub-Reddits Over Eleven Years. *Frontiers in Political Science*, 130.

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on Twitter. *Sage Open, 10*(2), 2158244020919447.

Thomas, M., & Joy, A. T. (2006). *Elements of information theory*. Wiley-Interscience.

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication, 33*(3), 460-480.

Thorson, K., Vraga, E., & Ekdale, B. (2010). Credibility in context: How uncivil online commentary affects news credibility. *Mass Communication and Society, 13*(3), 289-313.

Top, A. (2018). *500 Global Sites 2017*. https://www.alexa.com/topsites.

Tsai, W.-H. S., & Men, L. R. (2013). Motivations and antecedents of consumer engagement with brand pages on social networking sites. *Journal of Interactive Advertising, 13*(2), 76-87.

Tsvetkova, M., & Macy, M. (2015). The social contagion of antisocial behavior. *Sociological Science, 2*, 36-49.

Udris, R. (2017). Psychological and social factors as predictors of online and offline deviant behavior among Japanese adolescents. *Deviant behavior, 38*(7), 792-809.

Vale, L., & Fernandes, T. (2018). Social media and sports: driving fan engagement with football clubs on Facebook. *Journal of strategic marketing, 26*(1), 37-55.

Van't Riet, J., & Van Stekelenburg, A. (2021). The Effects of Political Incivility on Political Trust and Political Participation: A Meta-Analysis of Experimental Research. *Human Communication Research*.

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2014). Cursing in english on twitter. Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing,

Wang, Y., & Fesenmaier, D. R. (2004). Towards understanding members' general participation in and active contribution to an online travel community. *Tourism management, 25*(6), 709-722.

Warner, B. R., & Neville-Shepard, R. (2014). Echoes of a conspiracy: Birthers, truthers, and the cultivation of extremism. *Communication Quarterly, 62*(1), 1-17.

Wheeler, L. (1966). Toward a theory of behavioral contagion. *Psychological review, 73*(2), 179.

Wikipedia contributors. (2021a). *List of Most-Followed Facebook Pages*. Retrieved September 13 from https://en.wikipedia.org/w/index.php?titleList_of_most-followed_Facebook_pages&oldid1041704712

Wikipedia contributors. (2021b). *List of Most-Followed Twitter Accounts*. Retrieved September 13 from https://en.wikipedia.org/w/index.php?titleList_of_most-followed_Twitter_accounts&oldid1043863300

Wikipedia contributors. (2021c). *List of Most-Subscribed YouTube Channels*. Retrieved September 13 from https://en.wikipedia.org/w/index.php?titleList_of_most-subscribed_YouTube_channels&oldid1043463938

Wojcieszak, M., de Leeuw, S., Menchen-Trevino, E., Lee, S., Huang-Isherwood, K. M., & Weeks, B. (2021). No Polarization From Partisan News: Over-Time Evidence From Trace Data. *The International Journal of Press/Politics*, 19401612211047194.

Wojcieszak, M. E., & Mutz, D. C. (2009). Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of Communication, 59*(1), 40-56.

Wood, W. (2000). Attitude change: Persuasion and social influence. *Annual review of psychology, 51*(1), 539-570.

Wood, W., Christensen, P. N., Hebl, M. R., & Rothgerber, H. (1997). Conformity to sex-typed norms, affect, and the self-concept. *Journal of personality and social psychology, 73*(3), 523.

Ybarra, M. L., Mitchell, K. J., Wolak, J., & Finkelhor, D. (2006). Examining characteristics and associated distress related to Internet harassment: findings from the Second Youth Internet Safety Survey. *Pediatrics, 118*(4), e1169-e1177.

Zaheer, A., & Bell, G. G. (2005). Benefiting from network position: firm capabilities, structural holes, and performance. *Strategic Management Journal, 26*(9), 809-825.

Zelenkauskaite, A., & Niezgoda, B. (2017). "Stop Kremlin trolls:" Ideological trolling as calling out, rebuttal, and reactions on online news portal commenting. *First Monday*.

Zhao, S., Grasmuck, S., & Martin, J. (2008). Identity construction on Facebook: Digital empowerment in anchored relationships. *Computers in Human Behavior, 24*(5), 1816-1836.

APENDICES

Appendix A: Sub-Reddit Annotation for Chapter 2

Seven coders received initial training on the coding procedures, and each of them completed five iterative pilot coding exercises. Coders were instructed to code sub-Reddits based on the self-provided sub-Reddit description and the top popular posts in that sub-Reddit. More specifically, coders would go to the webpage of the sub-Reddit and read "about community" for an initial sense of the sub-Reddit. Then, they would read the top 20 posts on the first page. Although the coders only had access to the posts and comments from 2019; based on our knowledge about Reddit dynamics, sub-Reddits are rather stable in terms of discussion themes even though it is possible that sub-Reddits develop and change over time to some extent. If a sub-Reddit was active in 2019, the whole discussion in that sub-Reddit is identifiable via top posts and comments in 2019. If a sub-Reddit was "dead," the top posts from the last activity served the purpose of identifying the discussion topic. If 70% of posts were related to socio-political issues, the sub-Reddit was categorized as a political sub-Reddit. Then, the coders were instructed to judge by the top posts whether the sub-Reddit is homogeneous or heterogeneous, and its political leaning if the sub-Reddit is politically homogeneous. If less than 40% of the posts were related to socio-political issues, the sub-Reddit would be categorized as non-political, and coders would further categorize it based on its domain topic. The rest sub-Reddits would be categorized into the mixed group. Mixed sub-Reddits would be further categorized as politically homogeneous/heterogeneous and for their political leaning, as well as different non-political topics. Coding procedure took place from October 2019 to March 2020. The inter-coder reliability for sub-Reddit types of political/non-political/mixed (Fleiss's kappa=0.62), political homogeneous/heterogeneous (Fleiss's kappa=0.59), mixed homogeneous/heterogeneous

(Fleiss's kappa=0.85), political liberal/conservative (Fleiss's kappa=0.76), mixed liberal/conservative (Fleiss's kappa=0.59), non-political topics for mixed group (Fleiss's kappa=0.57), and non-political topics (Fleiss's kappa=0.83) were then calculated and met the fair requirement. These values of Fleiss' Kappa's are fair and acceptable, especially given that we had 7 coders, instructed to code communities into 10 categories, and because each community could be categorized into multiple categories. Coders then proceeded to individual coding, with 10% overlap in the coded content.

Appendix B: Sub-Reddit Examples for Chapter 2

Table 1

Example sub-Reddits for each category

Political			Mixed		
Homogeneous liberal	Homogeneous conservative	Heterogeneous	Homogeneous liberal	Homogeneous conservative	Heterogeneous
neoliberal	The_Donald	politics	europe	BlackPeopleTwitter	AskReddit
Libertarian	Conservative	news	NoStupidQuestions	india	unpopularopinion
nottheonion	Firearms	bestof	conspiracy	ar15	canada
changemyview	forwardsfromgrandma	CanadaPolitics	technology	CCW	wallstreetbets
weedstocks	Anarcho_Capitalism	Vaping	france	terriblefacebookmemes	Philippines
Futurology	AskThe_Donald	PoliticalDiscussion	australia	ScottishPeopleTwitter	MGTOW
MensRights	progun	ShitPoliticsSays	atheism	CombatFootage	legaladvice

Non-political									
Gaming	Entertainment	Sports	Health	Music	Science/ Technology	Pets/Animals	Lifestyple/Fashion	Others	Memes
FortNiteBR	funny	nba	Drugs	hiphopheads	pcmasterrace	aww	gonewild	todayilearned	dankmemes
gaming	movies	nfl	depression	Music	buildapc	cats	relationships	AskOuija	MemeEconomy

leagueoflegends	videos	soccer	opiates	popheads	Android	NatureIsFuckingLit	mildlyinteresting	Showerthoughts	me_irl
DestinyTheGame	anime	SquaredCircle	fatlogic	kpop	apple	dogs	CryptoCurrency	WTF	PewdiepieSubmissions
wow	marvelstudios	CFB	ADHD	techsupport	jailbreak	Aquariums	teenagersnew	interestingasfuck	AdviceAnimals
NintendoSwitch	DDLC	hockey	progresspics	Guitar	teslamotors	natureismetal	personalfinance	RoastMe	freefolk
RocketLeagueExchange	rupaulsdragrace	MMA	SuicideWatch	indieheads	hardwareswap	Eyebleach	trashy	mildlyinfuriating	hmmm

Appendix C: Incivility Classifier Building for Chapter 2

BERT is a deep transformer model pre-trained on huge amounts of unlabeled text using a word-masking training objective. By training on this objective with such a large amount of data, BERT builds a powerful language model which can then be fine-tuned to successfully complete specific language-related tasks, such as question answering and text classification. By pretraining the model on unlabeled text, the model is taught to understand the target language. Then by fine-tuning, the model learns how to complete certain specific tasks with its knowledge of the language. DistilBERT is a BERT model which has had its number of parameters reduced using a compression technique called model distillation, in which a smaller model is trained to replicate the output of a larger model using a teacher-student training technique (Sanh et al., 2019). The result is a smaller, faster model which retains much of the performance of the original model. According to Sanh et al. (2019), the size of a BERT model will be reduced by 40%, while 97% of its language understanding capabilities will be retained and it will also be 60% faster. Our models started with the respective base pretrained language models from HuggingFace's Transformer's package. The models were then further pretrained on a 3-million Reddit comment dataset sampled from the whole data with 100,000 training steps (Sun et al., 2019). A SoftMax layer for binary classification in the final hidden layer was utilized. Then the models were fine-tuned for four epochs on 5000 annotated comments with 10% of the data for training validation and 1000 coded comments were set aside for model testing (see Davidson et al., 2020 for details). Classification results for BERT and DistilBERT are in Table 1.

Our approach to deal with the data scale is to generate a large collection of labeled comments from our dataset by using the fine-tuned DistilBERT model as training data. Considering that the natural proportion of incivility in Reddit comments may lead to an imbalanced training set and potential of bias in models, an additional synthetic data for oversampling was generated using ADASYN (He et al., 2008). In total, we generated 5 million Reddit comments with labels tagged by our DistilBERT model. Logistic regression models were then uptrained on the synthetic data classified by DistilBERT model together with human coded sample data (see Davidson et al., 2020). Results are shown in Table 2.

Table 1

Classification Results from BERT and DistilBERT Models

	BERT	DistilBERT
precision	0.814	0.936
recall	0.759	0.702
accuracy	0.956	0.963

Table 2

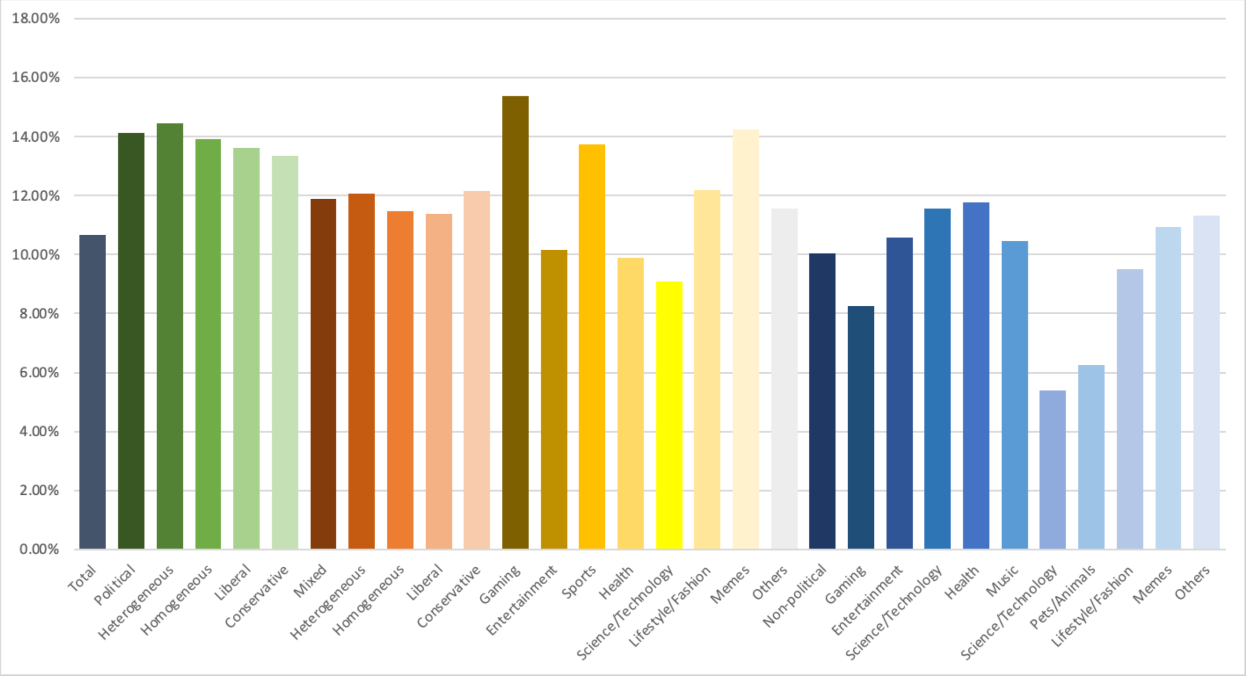
Classification Results for Different Training Data

Training Data	Precision	Recall	*F1*
Human coded data	1	0.173	0.295
Synthetic data	0.835	0.731	0.779

Appendix D: Average Incivility Proportion of All Categories for Chapter 2

Figure 1

Average Incivility Proportion for All Categories from 2008 to 2019



Note. The percentage shown in the graph is the proportion of incivility content over relative total content.

Table 1

Average Incivility Proportion for All Categories from 2008 to 2019

	Political	Mixed	Non-political	Total
Heterogeneous	14.46%	12.07%		
Homogeneous	13.92%	11.48%		
Liberal	13.61%	11.38%		
Conservative	13.35%	12.16%		
Gaming		15.36%	8.25%	

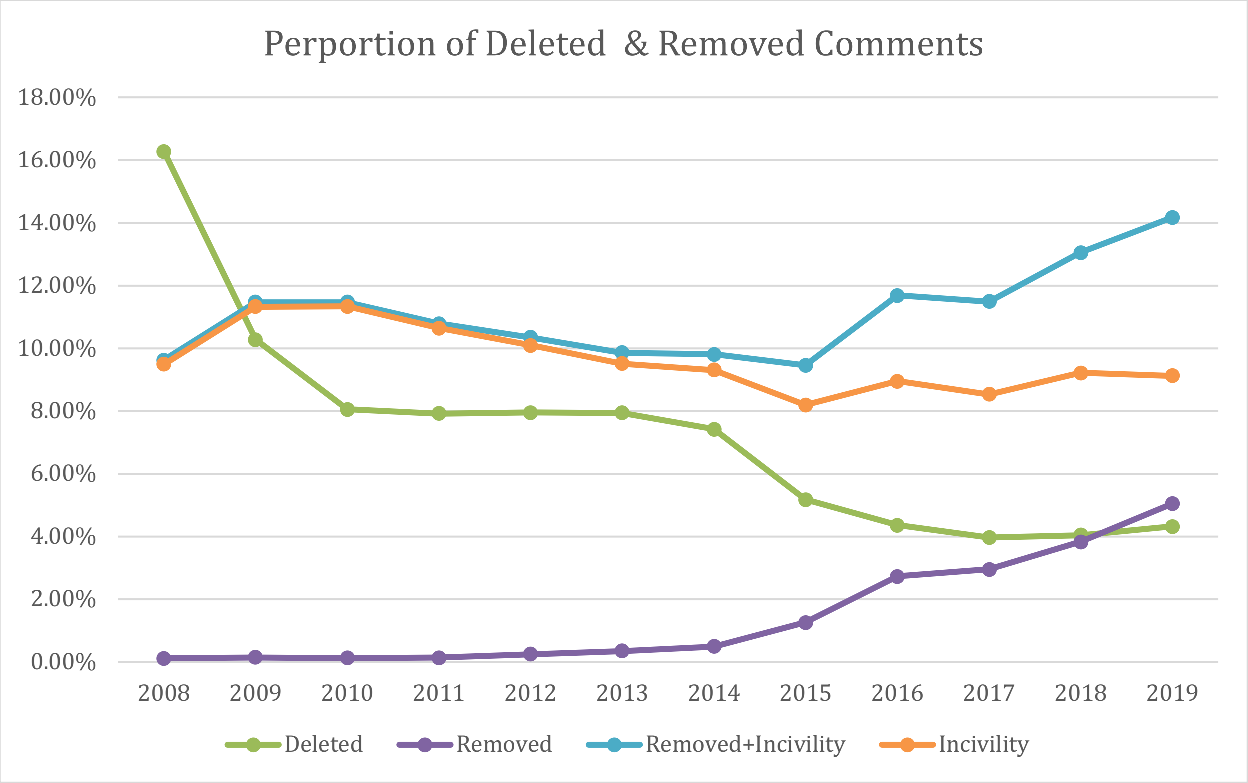
Entertainment		10.17%	10.57%	
Sports		13.75%	11.57%	
Health		9.89%	11.78%	
Music			10.46%	
Science/Technology		9.10%	5.40%	
Pets/Animals			6.24%	
Lifestyle/Fashion		12.20%	9.50%	
Memes		14.24%	10.95%	
Others		11.56%	11.32%	
Total	14.12%	11.89%	10.05%	10.68%

Appendix E: Additional Analysis for Deleted and Removed Comments for Chapter 2

To better understand the data bias caused by deleted and removed comments, we accessed the number of deleted and removed comments from 2008 to 2019. The difference between "delete" and "remove" is that the action of deleting is taken by user themselves, while the removed comments are taken down by the platform or moderators. There are multiple reasons to delete comments, for instance, regretting what had been said, the comments have been downvoted, or no one replied, which are not necessarily related to the use of incivility. On the other hand, the reason of removing comments is more likely to be a violation to Reddit or sub-reddit policy. The proportion of deleted and removed comments are showed in figure 1. The green line indicated the proportion of deleted comments, with the highest proportion as 16.28% in 2008 and decreased until to the lowest point as 3.97% in 2017. The purple line indicated the proportion of removed comments, which was under 1% before 2015, while after 2015, it increased every year and reached highest point in 2019 at 5.05%. It confirmed our observation of the 2015 anti-harassment policy that Reddit was actively engaged in establishing policy and moderating discussions. If we hypothetically consider all removed comments as uncivil comments, as shown in figure 1, the blue line, which is the total proportion of removed and uncivil comments, increased rapidly after 2015 comparing with the orange line, which represent the proportion of incivility (note that here the proportion is calculated with total number of comments including deleted and removed comments). It further suggested that the platform level intervention is effective in keeping the incivility level low on the platform but not necessarily in preventing people express uncivilly.

Figure 1

The Proportion of Deleted and Removed with Uncivil Comments from 2008 to 2019


Perportion of Deleted & Removed Comments
18.00%
16.00%
14.00%
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Deleted
Removed
Removed+Incivility
Incivility

Appendix F: Additional Analysis for Chapter 3

In addition to the likelihood of response (binary), we also ran mixed effect Poisson regression models as additional analysis to test the relationship between the frequency of response to a troll (excluding the first response), a count variable indicating how many times a community member responded to a specific troll, and the same set of independent variables (Table 3). The reason for excluding the first response is that most community members responded to a specific troll only once (69.35%), so that the logistic regression and the Poisson regression would have been highly similar without the exclusion. The Poisson regression results were presented as Model 3 and Model 4 in Table 3. The degree centrality, activeness of a community member, and total responses to trolls in general are found significantly related to the frequency of a community member's responses to trolls.