Information Processing and Management 58 (2021) 102660

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Developing a socio-computational approach to examine toxicity propagation and regulation in COVID-19 discourse on YouTube

Adewale Obadimu [a], Tuja Khaund [b], Esther Mead [b], Thomas Marcoux [b], Nitin Agarwal [b,*]

[a] LinkedIn Corporation
[b] Department of Information Science, University of Arkansas at Little Rock, Arkansas USA

ARTICLE INFO

Keywords:
Toxicity analysis
Social network analysis
Topic modeling
Pandemic
COVID-19
YouTube
Social media

ABSTRACT

As the novel coronavirus (COVID-19) continues to ravage the world at an unprecedented rate, formal recommendations from medical experts are becoming muffled by the avalanche of toxic content posted on social media platforms. This high level of toxic content prevents the dissemination of important and time-sensitive information and jeopardizes the sense of community that online social networks (OSNs) seek to cultivate. In this article, we present techniques to analyze toxic content and actors that propagated it on YouTube during the initial months after COVID-19 information was made public. Our dataset consists of 544 channels, 3,488 videos, 453,111 commenters, and 849,689 comments. We applied topic modeling based on Latent Dirichlet Allocation (LDA) to identify dominant topics and evolving trends within the comments on relevant videos. We conducted social network analysis (SNA) to detect influential commenters, and toxicity analysis to measure the health of the network. SNA allows us to identify the top toxic users in the network, which led to the creation of experiments simulating the impact of removal of these users on toxicity in the network. Through this work, we demonstrate not only how to identify toxic content related to COVID-19 on YouTube and the actors who propagated this toxicity, but also how social media companies and policy makers can use this work. This work is novel in that we devised a set of experiments in an attempt to show how if social media platforms eliminate certain toxic users, they can improve the overall health of the network by reducing the overall toxicity level.

1. Introduction

In recent years, we have witnessed an exponential growth in the amount of digital content that is being pushed on various social media platforms. Now, more than ever, online social networks (OSNs) have become a go-to place for obtaining news, information, and entertainment. Despite the myriad advantages of utilizing OSNs, a consensus is emerging suggesting the presence of an ever-growing population of malicious actors who utilize these networks to spread toxicity and harm others. These actors (hereafter referred to as toxic users) thrive on disrupting the norms of a given platform and causing emotional trauma to other users. To set a context for our work, we give an operational definition of toxicity as "the usage of rude, disrespectful, or unreasonable language that will likely provoke or make another user leave a discussion." (Obadimu, Mead, Hussain, et al., 2019). In this regard, toxicity analysis is different

* Corresponding author.
E-mail address: nxagarwal@ualr.edu (N. Agarwal).

https://doi.org/10.1016/j.ipm.2021.102660
Received 22 January 2021; Received in revised form 12 May 2021; Accepted 8 June 2021
Available online 10 June 2021

from sentiment analysis, which is the attempt to assign sentiment scores of positive, neutral, and negative to text data.

According to a report by the Pew Research Center, 41% of Americans have been the target of online harassment (NW et al., 2014). In a recent event that took the social media world by the storm, Quaden Bayles, a 9-year-old wanted to commit suicide after his colleagues bullied him because of his dwarfism (Srikanth, 2020). In the viral video, Quaden can be heard saying to his mother, "Give me a knife. I want to kill myself." In another unfortunate circumstance that happened in 2017, Gabriel Taye, an 8-year old boy ended his life after his classmates bullied, nudged, and poked him (*8-Year-Old Boy Commits Suicide After Being Bullied|PEOPLE.Com*, 2017). A journalistic account by The Guardian in 2013 discussed how a 12-year-old girl committed suicide after being targeted for cyberbullying (Haven, 2013). In 2012, Charlotte Dawson, who at one time hosted the "Next Top Model" TV program in Australia, committed suicide after being targeted with malicious online comments (Lee & Kim, 2015). Toxicity is a problem that is seriously affecting the dynamics and usefulness of online social interactions. Due to the growing concerns about the impact of online toxicity, many platforms are now taking several steps to curb this phenomenon. For instance, on YouTube, a user can choose to activate the safety mode to filter out offensive language.

OSNs like YouTube connect and stimulate interactions between people from all over the world. Unfortunately, those interactions are not always positive. YouTube provides a particularly fertile ground for toxic behavior, as most videos are accompanied by comment sections. Unlike other OSNs, there is no 'friending' or approval process to control who may comment on a video. Anyone with a user account can freely engage in this negative and harmful behavior, which often incites further negative actions, much like the spread of a disease. While toxic behavior is an unfortunate hallmark of online interaction, certain events can increase the prevalence of this behavior. The recent pandemic due to the novel coronavirus (COVID-19) is one such event.

The novel coronavirus, officially known as COVID-19, first emerged in Wuhan, China in December of 2019 and has spread to at least 167 countries and territories (*Coronavirus Disease (COVID-19) – World Health Organization*, 2019). As of July 29, 2020, there were 16,978,206 confirmed cases worldwide, with a death toll of 666,239 according to data compiled by the U.S.-based Johns Hopkins University. Since the health of the social media user is at risk of being negatively affected by the toxicity that is being interjected into social media platforms, and due to the potential power of the YouTube platform to be used as a vehicle for the propagation of toxicity, we have outlined the following research questions for this study: Is there toxicity in the online discourse around COVID-19 topics? If so, can this toxicity be identified and measured? Prior investigation into contagion and diffusion in online platforms have indicated that under the right circumstances, some forms of emotion can and does spread. In this study, we explore whether toxicity, like other forms of emotion, is contagious; or, in other words, does toxicity spread along with network ties? If that is the case, can such a phenomenon be observed and visualized systematically? Based on the answers to the aforementioned questions, can we develop methods to reduce toxicity? If so, how can such methods be evaluated? It is imperative that we develop techniques to identify toxicity and the commenters that attempt to propagate it. Hence, utilizing COVID-19 as a case study, our objectives in this paper are to identify and profile toxicity within the network, examine themes among toxic comments, investigate connections among toxic commenters, and simulate the removal of these commenters to discover the impact on the health of the network. In order to study these research questions, we outline a detailed methodology for analyzing a COVID-19 discourse dataset, and, subsequently, a technique for simulating how OSNs can improve the overall health of their networks by eliminating highly toxic users. We present the following contributions in this paper:

- Applied topic modeling based on LDA and Social Network Analysis to identify and visualize common themes among the toxic comments. Similar themes were found using both methods *(Section 3.2)*.
- Profiled common behavior patterns among toxic users. We discovered that commenters with similar toxic levels tend to stay together and replies to toxic comments often were at a similar toxicity level to the original comment *(Section 3.3.1)*.
- We identify information broker nodes that served as gatekeepers and controlled the flow of toxic information across various clusters within the network *(Section 3.3.2)*. This further led us into our simulation where we evaluate the overall health of the network after removal of such broker nodes and we observe that indeed the network recovers (*Section 4.2*).
- We observed how toxicity traversed within a network and affected non-toxic commenters (*Section 3.3.3*) as they come in contact with highly toxic commenters. High degree of segregation among commenters revealed that users mostly react to toxic comments with equal or high degree of toxicity. There were also groups found working collectively to form echo-chambers to amplify such toxic beliefs.
- Analysis of the content revealed potential inorganic or bot activity within the network, which is indicated by clusters that have a very tight formation and consist of identical comments being posted (*Section 3.3.4*). Such activity grew over the period of 4 months as the narratives transformed from minor insults to hate speech and profanity, etc. (*Section 3.3.5*).
- Used the discoveries from our analyses to perform simulations of the effect of removing these users on the average toxicity of the remaining network. We found that the greatest reduction of toxicity was achieved by removing users based on the computed toxicity score of their comments (*Section 4.4*).

The remainder of this paper continues as follows. First, we highlight extant literature that is germane to our work in the section on Literature Review. We then describe our data collection and processing in the Methodology section. Next, we delve into the analysis techniques applied to understand the data and our findings. We also present experimental simulations on the impact of removing toxic users in the network. Finally, we conclude with ideas for future work and a discussion of potential limitations of this work.

2. Literature review

To contextualize our work with broader literature, in this section, we discuss significant extant research that is germane to our

work. First, we discuss works relative to toxicity analysis. Next, we discuss works relative to contagion in social networks. Finally, we discuss works relative to topic modeling.

2.1. Toxicity analysis

Prior studies have shown that online users participate in toxic behaviors out of boredom (Varjas et al., 2010), to have fun (Shachaf & Hara, 2010), or to vent (Lee & Kim, 2015). A comprehensive examination of various forms of toxicity was conducted by Warner et al. (Warner & Hirschberg, 2012). Another study (Suler, 2004) suggests that toxic users have unique personality traits and motivation; however, Cheng et al. (Cheng et al., 2017) noted that given the right condition, anyone can exhibit toxic tendencies. Another study (Cheng et al., 2015) shows that toxic users become worse over time, in terms of the toxic content they post, and they are more likely to become intolerant of the community. One of the major obstacles in understanding toxic behavior is balancing freedom of expression with curtailing harmful content (Chen et al., 2012). Closer to our work is a study that analyzed toxicity in multiplayer online games (Märtens et al., 2015). The authors indicated that a competitive game might lead to an abuse of a communication channel.

Toxic behavior, also known as cyberbullying (Märtens et al., 2015), or online disinhibition (Suler, 2004), is bad behavior that violates social norms, inflicts misery, continues to cause harm after it occurs, and affects an entire community. Since toxic behavior has an offline impact, a deeper understanding of its properties is needed. Several researchers have tried to identify and suggest ways to mitigate hate speech in a community (Chen et al., 2012; Wulczyn et al., 2017). Using data collected via crowdsourcing, Wulczyn et al. (Wulczyn et al., 2017) employed machine learning techniques such as linear regression and multilayer perceptron to analyze personal attacks at scale. The authors concluded that the problem of identifying and categorizing hate speech at scale remains surprisingly difficult. Other researchers (Märtens et al., 2015) utilized Natural Language Processing techniques to detect the emergence of undesired and unintended behavior in online multiplayer games. Yin et al. (Yin et al., 2009) leveraged a set of regular expressions, n-grams, and supervised learning techniques to detect an abusive language.

Other research, (Sood et al., 2012) combined lexical and parser features to identify offensive language in YouTube comments. Davidson et al. (Davidson et al., 2017) presented a dataset with three kinds of comments: hate speech, offensive but non-hateful speech, and neither. Hosseini et al. (Hosseini et al., 2017) demonstrated the vulnerability of most state-of-the-art toxicity detection tools against adversarial inputs. After experimenting with a transfer learning approach, other researchers (Gröndahl et al., 2018) concluded that hate speech detection is largely independent of model architecture. The authors showed that results are mostly comparable among models but do not exceed the baselines. Although the aforementioned studies have explored various techniques of tackling toxicity within a social network, our study provide empirical evidence to not only identify the most toxic offenders and their behavior patterns but also to envision a network without them.

2.2. Contagion in social networks

In another study (Christakis & Fowler, 2013), authors claimed that obesity, like other forms of emotional contagion (Hatfield et al., 1993), may spread in social networks in a quantifiable and discernable pattern that depends on the nature of social ties. They further indicated that social distance appears to be more important than geographic distance within these networks since the weight gain of immediate neighbors does not have an effect on the chance of weight gain in egos. In another work, (Green et al., 2017), authors showed how modeling gun violence as an epidemic that spreads through social networks via interpersonal interactions can improve violence prevention strategies and policies. Their results suggest that an epidemiological approach, modeled on public health interventions developed for other epidemics, can provide valuable information and insights to help abate gun violence within US cities.

In our previous work (Obadimu, Mead, Hussain, et al., 2019), we outlined a methodology for identifying and scoring toxicity within the user-generated content posted on a specific OSN, YouTube. The analysis was able to shed light on how the existence and magnitude of toxicity changes when there are shifts in the narrative. Social network analysis techniques have also been used extensively in other studies. Extant literature has explored the 'spread' of obesity (Christakis & Fowler, 2007), 'alcohol consumption' (Rosenquist et al., 2010), 'smoking' (Christakis & Fowler, 2008), happiness (Fowler & Christakis, 2008), loneliness (Cacioppo et al., 2009), depression (Rosenquist et al., 2011), drug use (Mednick et al., 2010), and influenza (Christakis & Fowler, 2010).

2.3. Topic modelling

Previous studies (Obadimu, Mead, & Agarwal, 2019; Obadimu, Mead, Hussain, et al., 2019) have investigated toxic behavior on OSNs by using topic analysis and social network analysis (SNA) to discover connections among the types of harmful comments made. Closer to our work, authors (Chandrasekaran et al., 2020) examined the changing sentiments and trends surrounding various themes and topics surrounding COVID-19. They noted that government agencies, health care organizations, businesses, and leaders who are working to address the COVID-19 pandemic can be informed about the larger public opinion regarding the disease and the measures they have taken so that adaptations and corrective courses of action can be applied to prevent and control the spread of COVID-19. Other authors (S. S. Lee et al., 2011) present a recommendation system employing a modified latent Dirichlet allocation (LDA) model in which users and tags associated with an item are represented and clustered by topics, and the topic-based representation is combined with the item's timestamp to show time-based topic distribution. Another work (Mei et al., 2008) formally defined the problem of topic modeling with network structure. They proposed a novel solution to this problem, which regularizes a statistical topic model with a harmonic regularizer based on a graph structure in the data. Their technique combines topic modeling and social network analysis and leverages the power of both statistical topic models and discrete regularization.

3. Methodology

To understand the online toxicity surrounding the COVID-19 pandemic, we employed a methodology that consists of four components: 1) data crawling and processing, 2) toxicity analysis, 3) topic modeling, and 4) social network analysis. We leveraged the YouTube search Data APIv3[1] to extract channels, videos, and comments using the following keywords: coronavirus, corona, virus, COVID19, COVID, and outbreak. The time frame of our dataset spans the period from January 2020 through early May 2020. To reduce noise in extracted data, several data processing steps were performed including data formatting, data standardization and data normalization. We used the Python programming language to transform our data into a standard and normalized format by dropping the missing values, ensuring that every column is assigned to the correct data type, etc. Our dataset consisted of 544 channels, 3,488 videos, 453,111 commenters, and 849,689 comments (826,569 of which were unique comments). The comments were primarily from videos that were categorized as "News & Politics" (94%), followed by "Entertainment" (0.08%). We delve deeper into the dominant topics in a later section. Overall, however, in this work, we combine the techniques of topic modeling, toxicity analysis, and social network analysis to emphasize the usefulness of how this combination of techniques can be applied to social media data to achieve the potential objective of improving the overall health of an OSN.

3.1. Toxicity detection

The next step in our methodology was to compute toxicity scores for each comment in the dataset. To accomplish this, we leveraged a classification tool called Perspective API[2], which was developed by Google's Project Jigsaw and 'Counter Abuse Technology' teams, and has shown to be highly effective in other case studies (Pavlopoulos et al., 2019; Han & Tsvetkov, 2020). This model uses a Convolutional Neural Network (CNN) trained with word vector inputs to determine whether a comment could be perceived as "toxic" to a discussion. The API returns a probability score between 0 and 1, with higher values indicating a greater likelihood of the toxicity label being applied to the comment. Since toxicity scores are based on a probability score of 0 to 1, we focused our analysis on toxic comments that are 0.5 or greater in order to gain a deeper understanding of high toxic content. This step in our methodology was executed in order to accomplish the primary goal of this study, which was to show how the reduction of the toxicity level on an OSN can improve the overall health of the network. Table 1 is an excerpt of the resultant toxicity dataset where toxicity scores have been assigned to each comment. We have added "*" to mask profane words.

Fig. 1 shows the evolution of the mean toxicity score over the time frame of our analysis, January 1, 2020 through early May of 2020. Fig. 1 reveals that toxicity experienced more fluctuation over the months of January and February and had a tendency to smooth out over the months of March and April. This high fluctuation in the mean toxicity sores of our COVID-19 dataset was compared to another of our datasets that is a sample of comments made on more general videos for the same time period (January through early May), which consistently hovered around a mean toxicity score of 0.3 (based on a sample of 14,075,839 comments). Table 2 provides the categorical details of the comparison sample of general videos.

The World Health Organization announced on January 9 that a deadly coronavirus had emerged in Wuhan, China. After the announcement was made, there were many toxic comments targeted towards China for the cause of the pandemic. These events explain the high fluctuation in the average toxicity between the range of 0.57 to 0.85.

In the month of February, the president of the United States of America faced an impeachment trial, he was subsequently acquitted on February 5. There was also a watershed moment for the #MeToo movement when Harvey Weinstein got convicted on February 24. This is noticeable in our data due to the high average toxicity around mid-February. The coronavirus pandemic triggered a global recession as numerous countries went into lockdown in the month of March. Finally, the police-involved killings of George Floyd led to outrage and riots across the world to demand an end to police brutality and racial injustice. Due to some of these events, the average toxicity of comments around this period is high and stable. The next section discusses the content of these user comments as revealed via topic modeling.

3.2. Topic modeling

Perceiving the discussed topics in social media is imperative to detecting various topic facets and extracting their dynamics over time. Hence, we leveraged a topic model based on latent Dirichlet allocation (LDA) (Blei, 2003) to automatically discover evolving topics that are related to COVID-19. In LDA models, each document is composed of multiple topics. A topic is a collection of dominant keywords that are typical representatives of an entire corpus. In our case, each YouTube comment is considered a "document". Typically, with this technique, only one of the topics is dominant. The topics inferred by the LDA model may be viewed as operationalizing several important concepts related to framing, polysemy (or co-existence of meaning), and a relational approach to meaning. This technique allowed us to get samples of sentences from comments that most represented a given topic. We used NLTK and the spaCy English model to perform text pre-processing such as lemmatization, stemming, tokenization and the removal of stopwords. Table 3 provides the distribution of comments per month of our dataset time frame.

Figs. 6-8 show the evolution of the topic streams over the time frame of our dataset, January to early May of 2020. The tool used to create these topic streams figures was based on previous work by Marcoux et al. (Marcoux et al., 2020). The y-axis represents the

[1] https://developers.google.com/youtube/v3
[2] https://www.perspectiveapi.com/#/home

Table 1
Convenience sampling of four (4) toxic comments in our dataset.

Comment	Overall Toxicity
fu**in nerd made this sh*t lame	0.99
Crazy orange a**hole	0.98
Who the f*ck eat monkeys? You crazy f*ckers.	0.97
This is disgusting.drop a bomb on the entire country	0.97

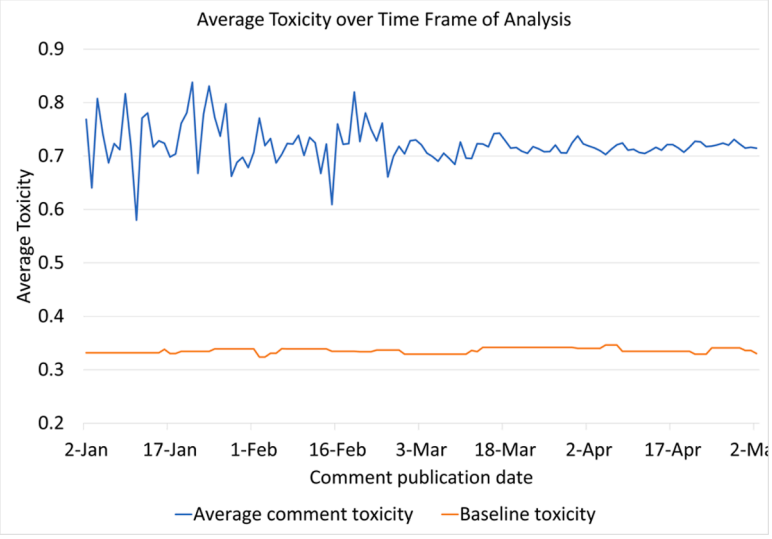

Fig. 1. Trend line comparison for average toxicity score among comments made on COVID-19 YouTube videos (blue) and those made on more general videos (orange) over the time frame of our analysis, January to early May of 2020.

Table 2
Distribution of video and comment counts per category of sample dataset of general YouTube videos used for comparison with the COVID-19 videos comments dataset.

Category	Comment Count	Video Count
News & Politics	11,056,419	40,702
Entertainment	2,431,338	6,078
People & Blogs	327,702	603
Other	260,380	2,645
Total	14,075,839	50,028

distribution of each topic within the dataset. Fig. 6 reveals the evolution of the topics when a window of 1,000 comments is used. This view shows topic 18 as being consistently dominant, which consists of the following keywords: trump, stupid, lies, democrats, fool, ignorant, shame, liars, ppl. Topic 11 was the second-most dominant topic in February, consisting of the following keywords: liar, crap, garbage, corrupt, sense, right, trust, seriously, facts. Topic 0 assumed and consistently held the second-most dominant position, however, beginning in March, consisting of the keywords: hate, white, racist, black, wtf, fox, land. Topic 11 still had a continual dominant presence in the dataset, however, as did topic 3, which consists of the keywords: 'biden', 'son', 'joe', 'clown', 'old', 'orange', 'like', 'office'. Table 4 shows the list of the topics and their corresponding words distribution.

Fig. 7 reveals the evolution of the topics when a window of 10,000 comments is used. This view allows for a clearer distinction between the dominance of topics over time. The figure more clearly shows that topic 18 was continually dominant, followed by topic 0 and topic 11. Topic 5 is revealed to have prominence, at times shifting back and forth between the fourth- and third-most dominant position. Topic 5 consists of the following keywords: idiot, con, fu**, gobierno, virus, omg, bin. The toxicity within topic 5 is evident by some of the keywords.

Fig. 8 reveals a better view of the evolution of the topics after a moving average was applied to the data. The figure is showing the month of April because it comprises 86.7% of our dataset. This view also reveals that Topic 18 consistently had the highest distribution. Topic 15 consistently had the second highest distribution except for a few days within the last week of April where topic 1 became more prominent. Topic 15 consists of the following keywords: vote, head, state, human, numbers, get, inject, nation, protesting, science; whereas, topic 1 consists of: video, fu**ing, dam*, look, sorry, wow, comments, really. The toxicity within topic 1 is evident by some of

Table 3
The distribution of comments per month of our COVID-19 YouTube videos comments dataset for time frame, January to early May of 2020.

January	185
February	2,075
March	86,755
April	760,674
May	27,393
Total	877,082*

*Data set used for generating topic streams included some additional data for the month of May.

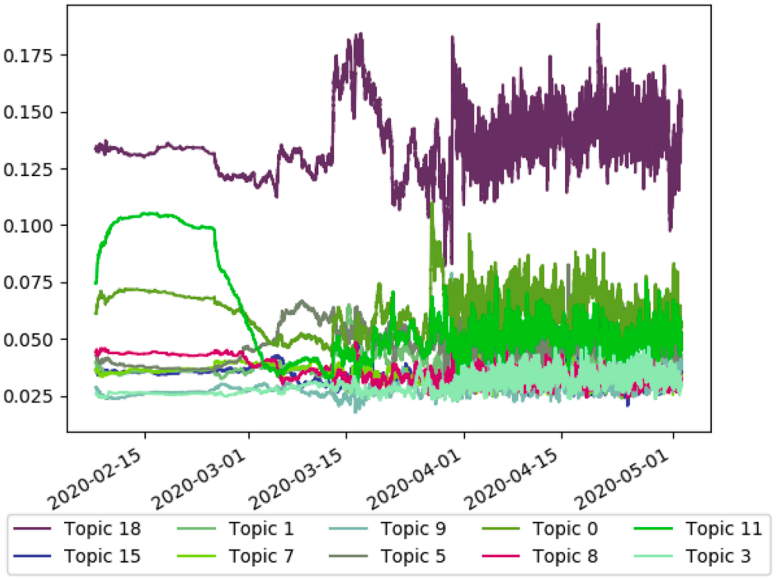

Fig. 6. Evolution of COVID-19 related topic streams from April to early May of 2020 using a window of 1,000 comments.

Table 4
Topics and their corresponding words distribution.

Topic	Words
8	People, one, even, would, like, know, time, new, said, every
16	People, get, want, like, going, back, need, take, think, would
7	people, covid, stupid, virus, Obama, stupidity, many, flu, kill, deaths
12	China, world, chinese, virus, ccp, communist, usa, Donald, government, blame
0	Das, hate, white, racist, black, del, wtf, este, fox, land
18	Trump, stupid, lies, democrats, fool, ignorant, shame, says, liars, ppl
5	Que, por, idiot, con, f*ck, una, gobierno, virus, omg, bin
1	Like, lol, video, f*cking, d*mn, look, sorry, wow, comments, really
9	News, fake, president, cnn, shut, guys, morons, talking, mouth, follow
17	Die, man, corona, virus, let, gonna, war, guess, also, guns
2	God, sh*t, hell, man, love, good, name, fools, believe, makes
13	America, country, president, states, death, pathetic, American, joke, worst, great
11	Liar, yeah, crap, garbage, corrupt, sense, right, trust, seriously, facts
4	Dumb, gates, bill, hat, evil, guy, vaccine, esse, nada, leader
14	Com, moron, watch, https, lie, lying, youtube, bunch, fear, www
15	Vote, head, state, human, numbers, get, inject, nation, protesting, science
19	Media, Bullshit, came, big, Australia, called, funny, try, said, liberal
6	Les, idiots, americans, mask, supporters, wear, dans, wish, check, hand
3	Los, biden, son, joe, clown, old, orange, like, per, office
10	Please, brain, use, comment, sch, system, sucks, drink, low, actually

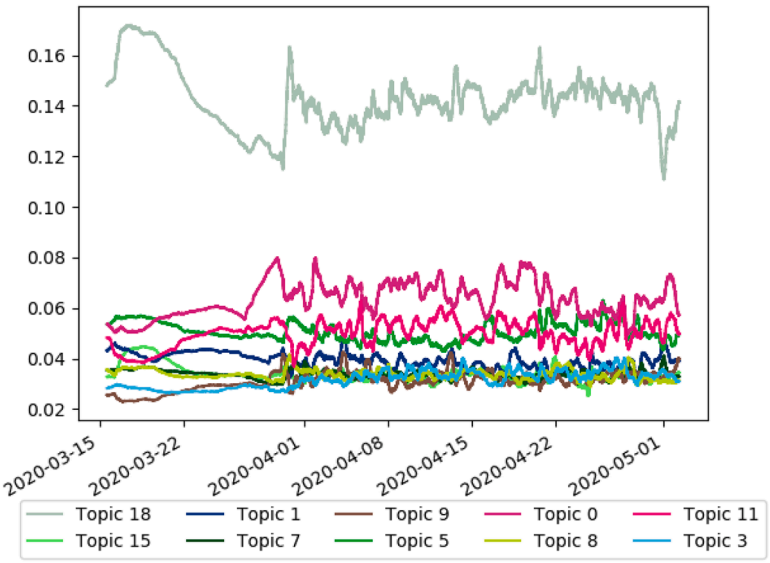

Fig. 7. Evolution of COVID-19 related topic streams from April to early May of 2020 using a window of 10,000 comments.

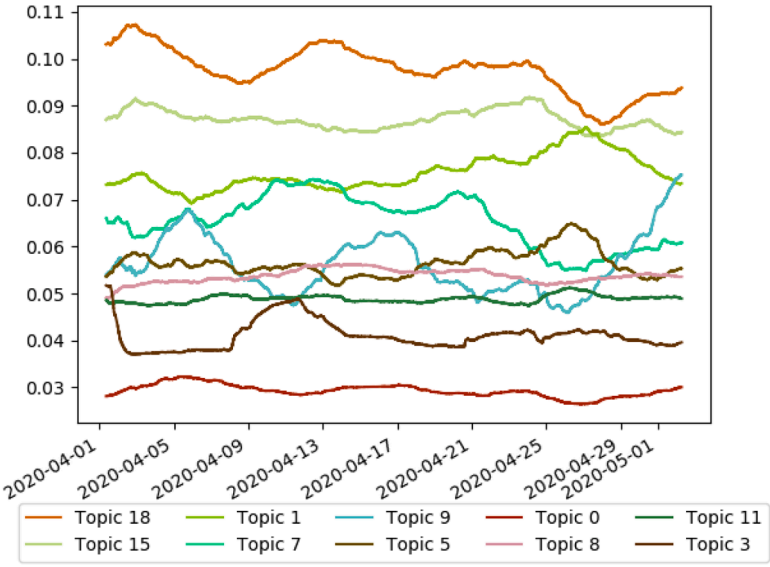

Fig. 8. Moving average of the evolution of COVID-19 related topic streams from April to early May of 2020.

the keywords. The next section details the application of social network analysis to our COVID-19 YouTube video comments dataset.

3.3. Social network analysis

We sought to analyze networks that profile common behavior patterns among toxic users (Cheng et al., 2015), including leaving comments on multiple videos within a channel, replying to other commenters with toxic content, and repeating/duplicating comments across videos. The results of our analysis are detailed in this section. Social network analysis (SNA) was performed on 145 videos that were posted in 110 YouTube channels, with a focus on the 32,107 unique users that left comments on these videos. In the accompanying network visualizations, the nodes are colored based on their toxicity scores (0.5 to 1) with the color ranging from blue (lowest toxicity), green, then yellow, then to red (highest toxicity). As the focus was on highly toxic people and the contagion of toxicity, we chose to remove the low toxicity scores (below 0.5). One other reason for excluding low toxic users was to reduce the complexity and computational cost of data processing. Where appropriate, we discuss important centrality measures. SNA allows us to identify the top toxic users in our network, which helped to inform experiments simulating the impact of the removal of these users.

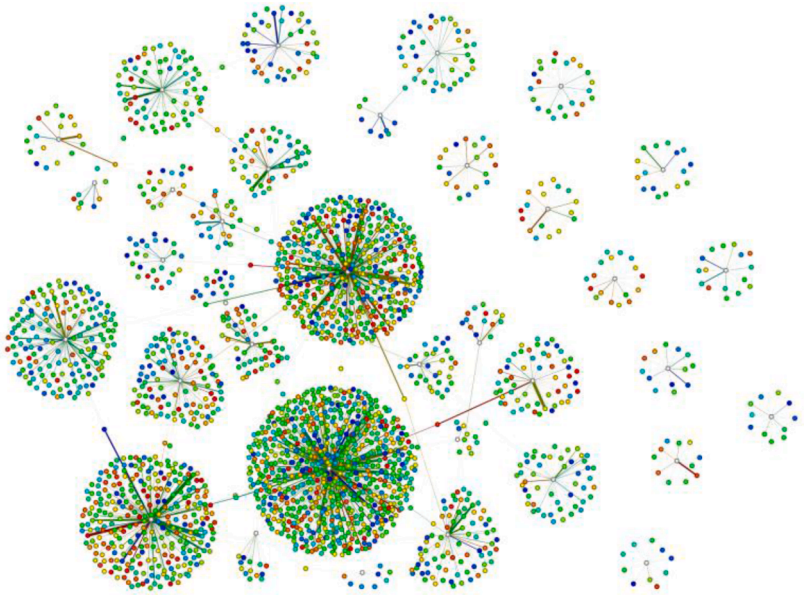

Fig. 9. Channel-commenter network. Colors denote toxicity scores from lowest 0.5 (blue) to highest 1 (red).

3.3.1. Channel-commenter network

Toxic and abusive commenters rarely limit their behavior to just one video. As YouTube channels typically contain several videos with similar content or subject matter, we've found that toxic users often spread their comments across multiple videos in one channel. In order to visually identify the toxicity levels, we first analyzed the highest-level view of the channel-commenter clusters. This helps us understand how an overall toxic network reflects as far as formation of clusters or connections. Is the overall network a toxic one or does it consist of smaller individual clusters? Fig. 9 illustrates this first highest-level overview of the network to reveal the toxicity levels that exist within the channel-commenter clusters that were detected.

We modified the network by including users who commented more than once during the January 1, 2020 to April 30, 2020 time frame[3]. In order to optimize the potential information inferences from the visualization of the network, we removed smaller components such as dyads, triads and any channel cluster with ten or less commenters. We noticed that the majority of the remaining channels consist of moderate to high toxic discourse. There exist several independent clusters of commenters as well as shared clusters where commenters posted toxic comments across multiple channels. This shows that some users are not confined to a single source of information but tend to spread their content across different outlets (channels).

3.3.2. Understanding information brokers and gatekeepers

Toxic comments do not often exist in isolation. Instead, some videos attract multiple toxic users. For such toxic users, we wanted to know who are the information brokers or agents that bridge multiple communities within this toxic network. The co-commenter network sought to examine the extent to which the users in our dataset commented on the same videos and also served as gatekeepers of information. In the network depicted in Fig. 10, each node represents a user, and the edge represents a video in common. An initial set of 2,453 users commented on the same video with at least one other user. We removed components such as dyads, triads and any cluster with ten or less commenters and modified the network to consist of 2,263 users. The resulting network shows the biggest components within the network.

We identified a series of smaller clusters that shared one common commenter in the largest connected component consisting of 2,057 nodes and 638,744 edges. These commenters served as gatekeepers by forming bridges between the smaller clusters enabling them to spread information quickly. Gatekeepers can select, channel, shape, manipulate, and delete information from the network solely based on their position. If any of these commenters are removed, the network structure will disintegrate. These users are also very toxic in nature. This figure also ties into our proposed simulation experiments (Section 4.2) for showing how removing highly toxic users can remove some of these bridges and increase the health of the network.

3.3.3. Understanding how toxicity can be contagious

Toxic content is also often present as replies to video comments (Almerekhi et al., 2020). Fig. 11 shows our commenter-reply network, where we profiled the patterns among the 7,885 replies to comments present in our dataset. This was created as a directed graph to demonstrate the flow of conversation and understand how toxicity spreads from nodes to nodes and becomes

[3] May 2020 data was not included in the social network analysis due to its small size.

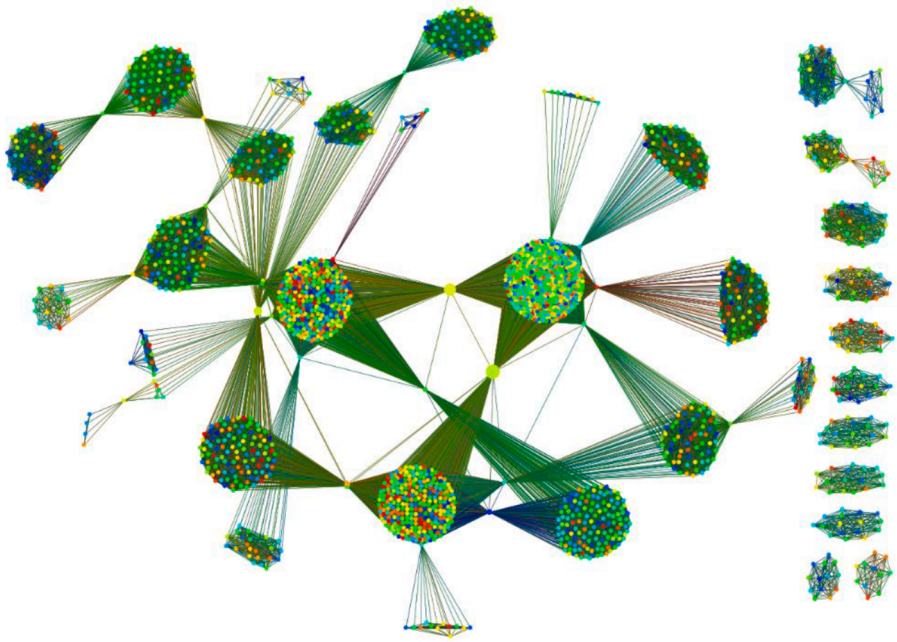

Fig. 10. Co-commenter shared video network. Colors denote toxicity scores from lowest 0.5 (blue) to highest 1 (red).

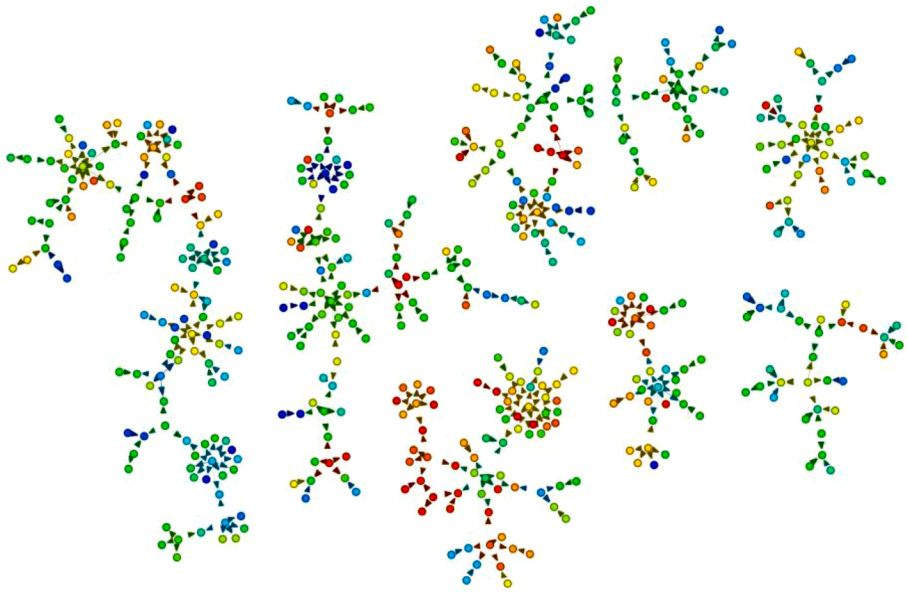

Fig. 11. Commenter-reply directed network clusters depicting segregation among commenters based on toxicity. Colors denote toxicity scores from lowest 0.5 (blue) to highest 1 (red).

contagious. Each edge represents a reply relation between a commenter and the replier with the arrow pointing towards the replier. We modified the network by removing components such as dyads, triads and any channel cluster with less than 30 commenters in order to make the network legible and focus on the larger clusters within the network.

We observed a high degree of segregation among commenters based on toxicity. The red or blue colored nodes are collocated most of the time suggesting that the replies to toxic comments had a similar level of toxicity. In other words, highly toxic commenters are more likely to form groups with commenters with high toxicity and low toxic commenters tend to cluster with commenters who are less toxic. Even though the network components are quite isolated from one another, various connections existed within these individual components. Modularity is a network measure designed to evaluate the strength of division of a network into groups or clusters. So, we

applied a modularity-based community detection algorithm (Newman & Girvan, 2004) to detect groups within the commenter-reply network components (see Fig. 12).

We obtained a high modularity score of 91.34%, which means that there exist dense connections between the nodes within clusters, but sparse connections between nodes in a different cluster. In other words, members of the same cluster are strongly connected whereas intra-cluster relationships are weak. Figs. 11 and 12 connect to show that the network is not only segregated but there are also groups that exist who are collectively forming "echo-chambers" to amplify toxic beliefs reinforced by communication inside a closed system. We also observed multiple different sub-clusters within the same component (Fig. 12, circled areas), which indicates, "network homophily" or assortativity based on the toxicity scores. Homophily simply means 'birds of a feather flock together'. In network science, homophily states that similar nodes have a greater tendency to group together than do dissimilar ones based on node attributes (toxicity).

3.3.4. Indications of inorganic activity within the network

The final toxic behavior pattern we studied was the tendency of toxic users to duplicate and repeat the same comment on multiple videos. Some of the popular traits of inorganic accounts is repetition and amplification based on existing literature (Qi et al., 2018). We wanted to study if any such bot activity existed in these commenter communities. We constructed a co-commenter shared comment network to highlight users posting identical comments multiple times and mapped these comments to the various commenter communities. We found 117 duplicate comments shared among 213 commenters. We modified the network by removing components such as dyads, triads and clusters with less than 5 commenters. These filters eliminated discussions from January 2020 completely, as the behavior was more common during the remaining months represented in our dataset (February to April of 2020). The highly toxic communities that are suspected of inorganic behaviors are highlighted in Fig. 13, along with the actual content of some of the most repeated comments.

Toxic users formed groups accusing the President of the US as being a "Racist", "Idiot", "Moron", etc., while a few highly toxic users protested against China with hashtags such as "#chinesevirus". The overall discussion focused on the novel coronavirus but there also existed non-relevant content pushed by certain commenters such as "f*** liverpool", "Russian troll farm conspiracy theories…", etc. These clusters of highly toxic users are suspected as being indicative of inorganic (or bot) activity due to not only the repetition of exact content, but also due to their very tight formation. Although this figure alone is not conclusive evidence of bot activity, it is enough of an indication to lead us to want to further explore this network in future work to determine some more conclusive findings with regard to this issue of suspected inorganic behavior.

3.3.5. Growth of inorganic activity through toxic comments over time

We subsequently drilled down to analyze the dominant comments for each of these three months separately (Figs. 14, 15, 16) for a better indication of the structure and content of some of the most toxic clusters per month. This helped us study how the network toxicity evolved over time and if the number of toxic comments increased along with it. These three figures in succession reveal how the toxicity within the network has evolved over time, in terms of score, content, and structure. The prominent clusters from the previous network (Fig. 13) remain prominent in the following networks as well. The narratives morphed over time and the topic trends

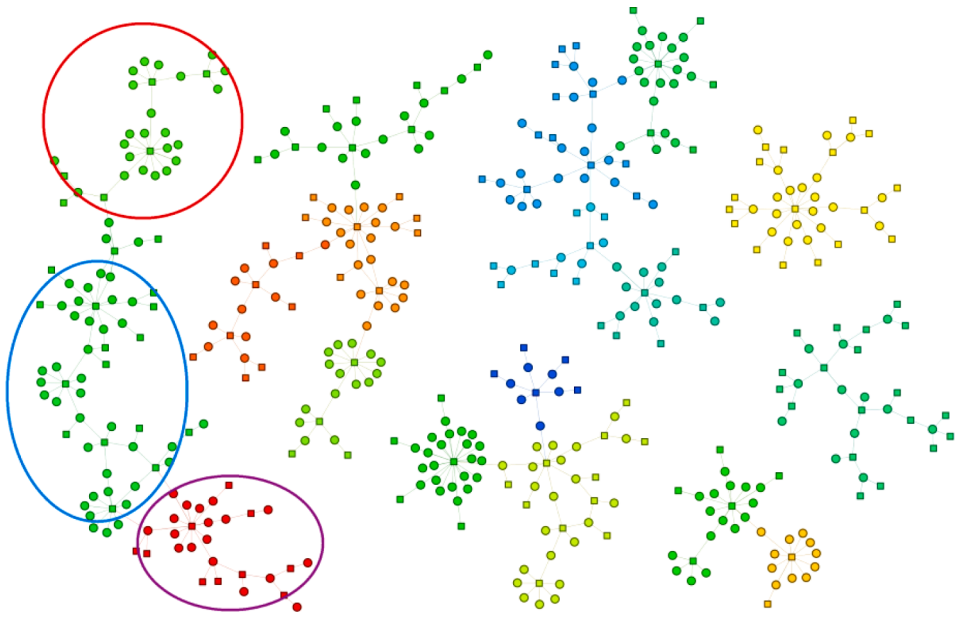

Fig. 12. Commenter-Reply Network clusters depicting high echo-chamberness. Square nodes are commenters and circular nodes are repliers. Colors denote the various clusters based on their modularity class.

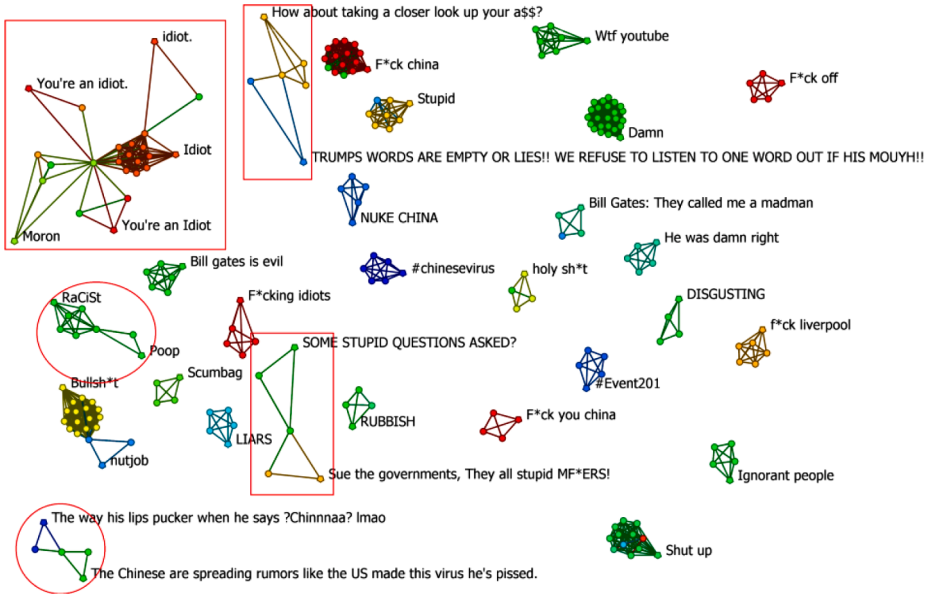

Fig. 13. Co-commenter Network based on shared comments from February to April of 2020. Colors denote toxicity scores from lowest 0.5 (blue) to highest 1 (red).

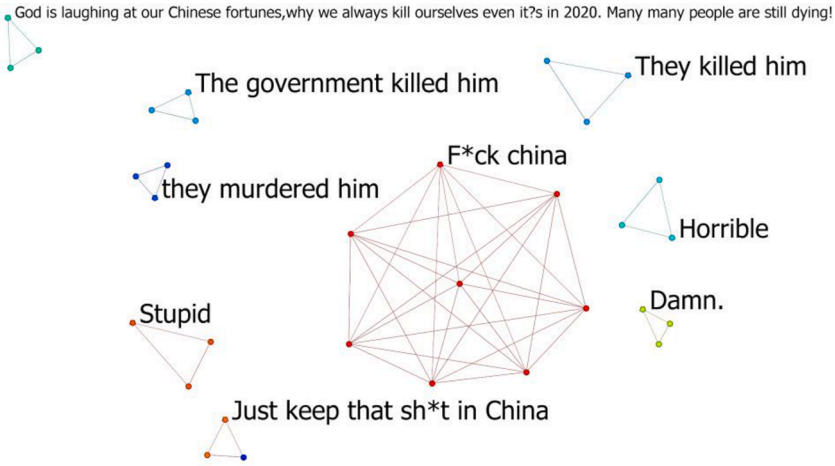

Fig. 14. Co-commenter Network based on shared comments in February of 2020. Colors denote toxicity scores from lowest 0.5 (blue) to highest 1 (red).

discovered from the topic modeling analysis were also found among the comments most duplicated and spread among the videos.

February: The network is quite small since it was still in the early stages of the global pandemic (Fig. 14). The themes of discussion for this month were mostly anti-Trump and anti-China. We identified a few clusters that accuse the government in particular of being responsible for the deaths of civilians and also China was mentioned since news outlets reported the source of COVID-19 could be from Wuhan, China.

March: As time progressed, we witnessed a growth in the number of comments as well as shared comments among toxic commenters (Fig. 15). The core narrative is still anti-Trump and anti-China, but these messages have now transformed into hate speech with commenters posting comments such as "NUKE China", "I hate China", "#chinesevirus", etc. Several commenters called out President Trump as being a "racist", "clown", "moron", "Liar" etc.

April: The toxic discourse on the various videos increased tremendously during April (Fig. 16). As toxic commenters continued to target and verbally abuse China with new comments such as "Just Nuke China already" and "NUKE CHINA NOW", within the cluster that initially posted "NUKE China", there originated new anti-Democratic party clusters that posted comments such as "*Why do Democ RATS always look so dirty, greasy, and smelly?*", "*Democrats credibility is dropping faster than Bill Clinton's pants on Epstein's Lolita Island!!!!*", etc. The final prominent cluster criticized President Trump with comments such as "Trump has blood on

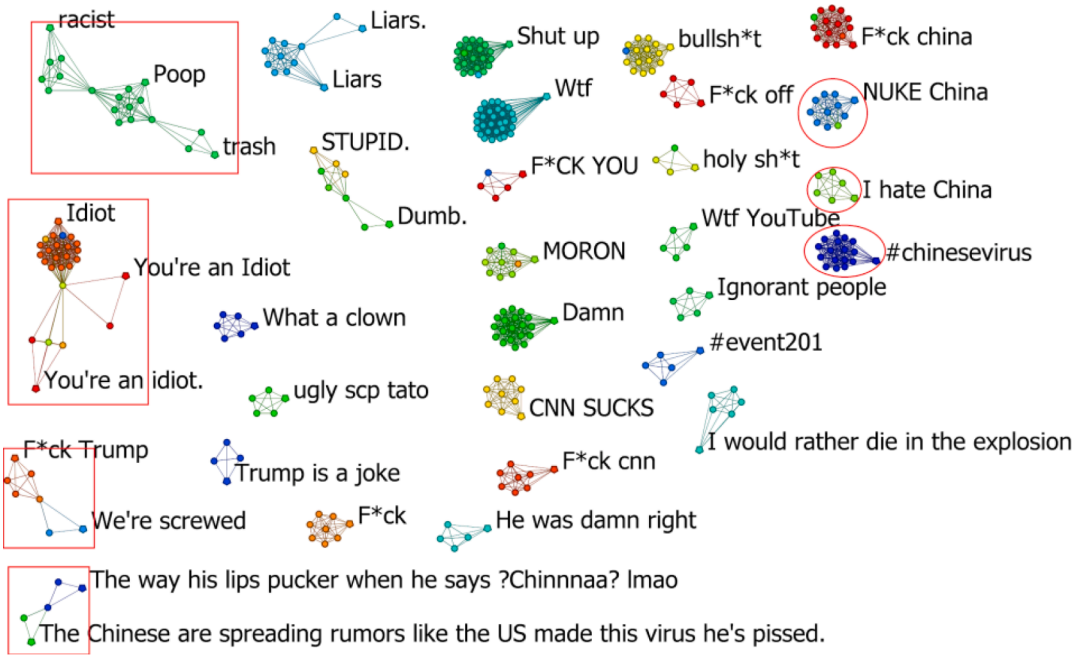

Fig. 15. Co-commenter Network based on shared comments in March of 2020. Colors denote toxicity scores from lowest 0.5 (blue) to highest 1 (red).

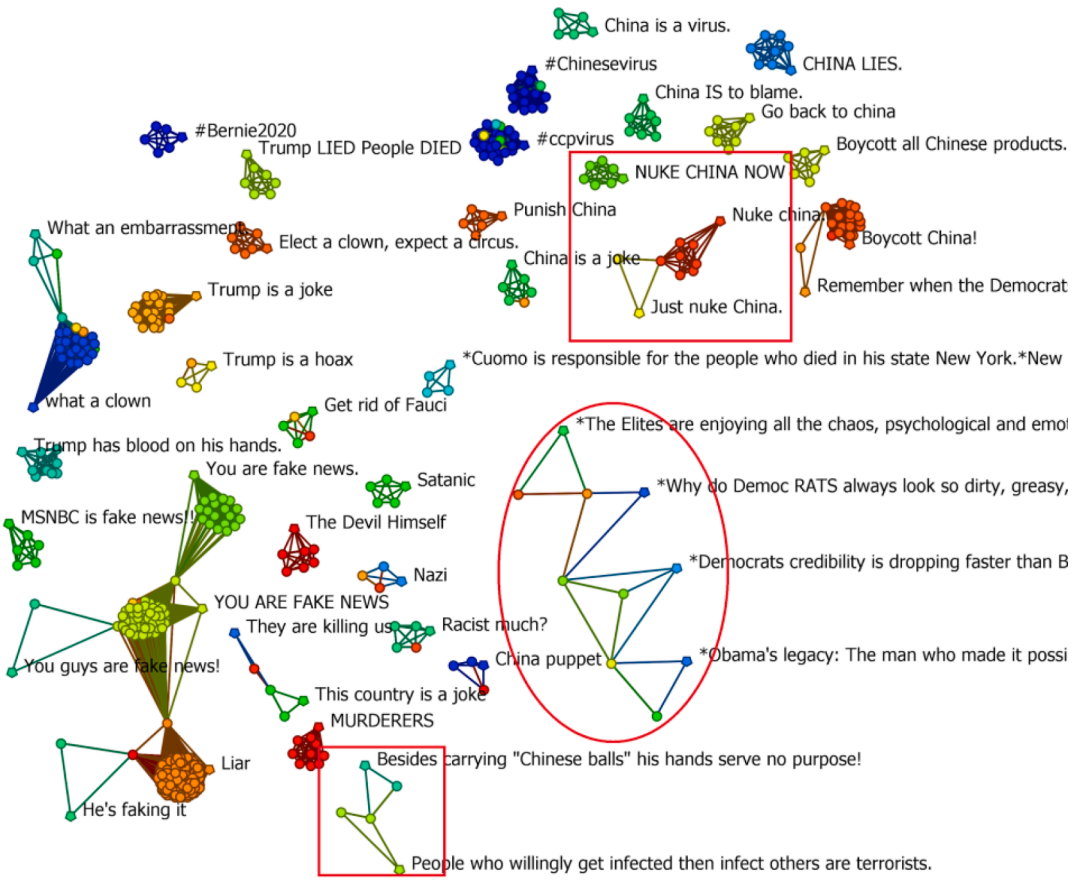

Fig. 16. Co-commenter Network based on shared comments in April of 2020. Colors denote toxicity scores from lowest 0.5 (blue) to highest 1 (red).

his hands.", "Trump LIED", "Besides carrying "Chinese b***s" his hands serve no purpose!", etc.

As the dataset developed over the course of four months, our analysis was successfully able to map identical comments to their commenter communities to hint on the presence of inorganic activity. We identified single word comments as well as phrases and sentences that were duplicated and amplified by multiple commenters. We also encountered more toxic comments as time progressed. However, this is still not enough to conclude the actual presence of bots. Therefore, future experiments need to be conducted to compare these patterns to existing bot detection methods. The next section discusses our toxicity reduction simulations.

4. Experimental simulations

In order to show how the administrators of OSNs can potentially improve the overall health of their networks, we conducted one three-fold experiment. Starting with the co-commenter shared video network shown in Fig. 10, we performed a cascading reduction of nodes based on certain metrics for each experimental section, rather than starting from scratch for three separate experiments. Our experiment attempted to use different methods to remove the most toxic users in the network. The following sections reveal the effects of the simulated removal of toxic nodes and overall improvement of the network health.

4.1. Identifying prominent toxic users

In this experiment, we use the network shown in Fig. 10 to identify bridge nodes, which are based on betweenness centrality (Newman & Girvan, 2004). Betweenness centrality measures help identify prominent users or nodes in a network that serve as bridges to two or more diverse clusters. The removal of bridge nodes will lead to the disintegration of the network structure. Since bridge nodes serve as a medium for toxicity propagation, it is best to first remove them to analyze the decrease in overall toxicity. For a nodeset of 2, 263 commenters, the mean toxicity is 0.72249617, which indicates that the entire network is quite toxic. Table 5 lists the top 10 commenters with high betweenness centrality and also reveals that those commenters are associated with very high toxicity levels. Next, we discuss how removing these nodes can potentially improve the overall health of the network.

4.2. Removing users based on betweenness centrality

In graph theory, betweenness centrality is a measure of centrality in a graph based on shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex. Betweenness centrality (Perez & Germon, 2016) can be measured as follows:

$$g(v) = \Sigma_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

This metric represents the degree to which nodes stand between each other. A node with higher betweenness centrality would have more control over the network, because more information will pass through that node. So, we began this experiment by removing the top ten nodes with the highest betweenness centrality value as listed in Table 5 and observed the overall toxicity score of the remaining network.

We observed that the initial set of 12 components is now increased to 19 which suggest that the network has indeed disintegrated. However, the network structure is not dissolved into isolates which suggest that even by removing top bridge nodes; the network was still able to sustain its structure. The network simply broke down into smaller clusters (see Fig. 17) with other low ranked nodes taking the place of the nodes that were removed. The mean toxicity of the resulting network of 2,243 commenters is slightly reduced to 0.720981759. The toxicity score had a 0.21% reduction from the previous score of 0.72249617.

Table 5
Top ten commenters based on betweenness centrality and their corresponding toxicity scores.

Commenter	Betweenness centrality	toxicity
UC1wmr8wGzfr6rV M4DzsiKA	0.262	0.79589
UCOASO9ytuYoaDsV3Bq6qdqw	0.224	0.83773
UC1QrrsMb71eI3JU l3-BQ3g	0.169	0.83232
UChCJz3Ab5EYgn-DkX-kaNcw	0.115	0.50474
UCCbIEMAkUS4qyysGc0s AKg	0.113	0.7355
UClgY3UBYNNWIVZJtzfLpwZQ	0.113	0.90138
UC-PgArmXAMgnM5JwH5qL8aQ	0.101	0.68622
UCN209zxAF 6fIOyRyBoVlXw	0.081	0.87247
UCmqsP0F8JrlGTjee8bI3rsw	0.042	0.96268
UCQ0naYA2mFWtI6BLi311RXA	0.040	0.83785

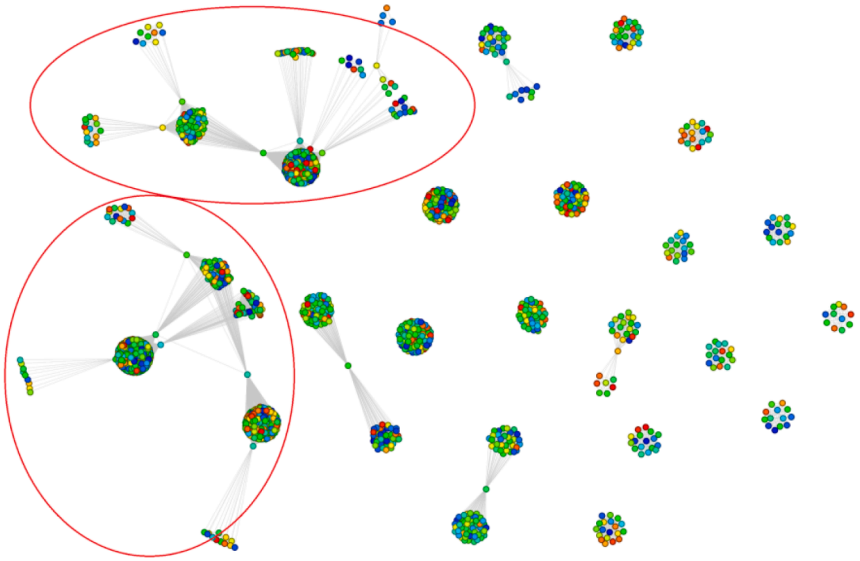

Fig. 17. Co-commenter shared video network with top betweenness centrality nodes removed. Colors denote toxicity scores from lowest 0.5 (blue) to highest 1 (red).

4.3. Removing users based on PageRank centrality

PageRank (PR) is an algorithm used by Google Search to rank web pages in their search engine results. PageRank is a way of measuring the importance of website pages. According to Google:

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites (*Google,* 2011).

PageRank satisfies the following equation:

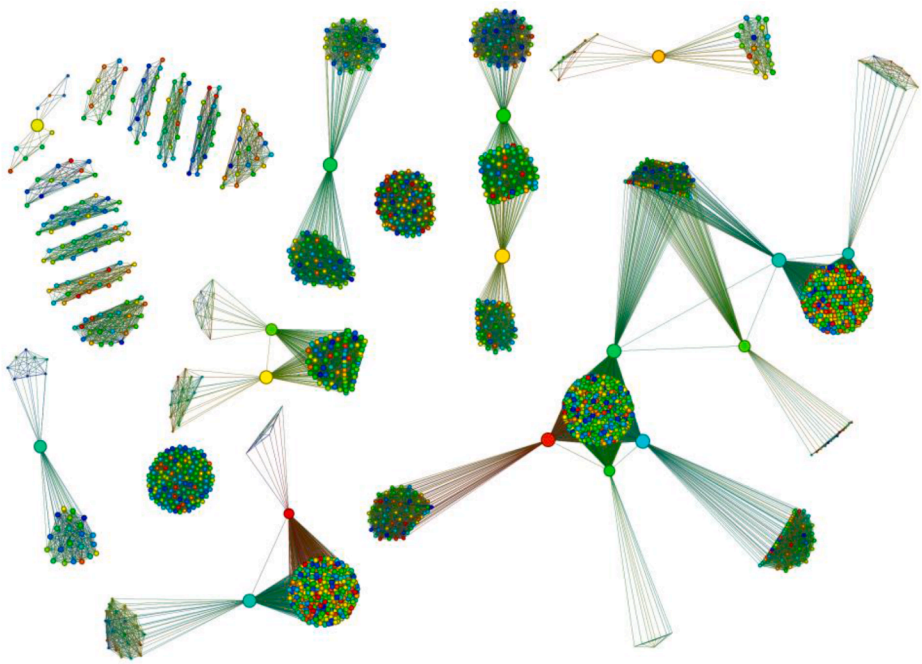

Fig. 18. Co-commenter shared video network with top ten nodes with high PageRank removed. Colors denote toxicity scores from lowest 0.5 (blue) to highest 1 (red).

$$x_i = \alpha \sum_j a_{ji} \frac{x_j}{L(j)} + \frac{1 - \alpha}{N},$$

where $L(j) = \sum_i a_{ji}$ is the number of neighbors of node j (or number of outbound links in a directed graph).

Since removing bridge nodes did not reduce the overall toxicity of the network tremendously, we used the PageRank (Heidemann et al., 2010) centrality measure to identify the important nodes in the network. We then removed the top nodes to analyze the degree to which toxicity decreased in the overall network. This measure helps us identify important nodes based on the importance of its neighbors. We observed that the network has now disintegrated to 20 larger components increasing from 12 that were initially identified. The network toxicity reduced by 0.02% which is significantly low. However, the toxicity score of 0.722317191 is still comparatively higher than our previous betweenness centrality experiment that generated a score of 0.720981759. This shows that even though we removed important nodes, the overall toxicity of the network did not change as opposed to removing bridge nodes despite being disintegrated into a larger number of components. The resultant network structure is shown in Fig. 18.

4.4. Removing users based on toxicity score

Although our previous experimental steps disintegrated the network, the overall change in toxicity was minimal. Our next step was to evaluate the impact of removing highly toxic users from the network. We chose a toxicity threshold greater than 0.8 since it captures most comments that are highly toxic. Seven hundred nodes were removed, and the remaining network consisted of 1,543 nodes (users) and 298,458 links. The nodes are colored based on toxicity and sized based on PageRank centrality. The network toxicity reduced by 11.15% with a score of 0.641927323 which is a significantly high level of toxicity reduction. This indicates that the overall health of the network improved after removal of highly toxic users. This helps us identify toxic users and report them to the admins of the social media platforms. These accounts can then be removed thereby reducing the spread of toxic content. The resulting network is shown in Fig. 19 and Table 6 summarized the percentage reductions in toxicity scores using the various strategies in our experimental simulations.

5. Conclusions and future work

In this work, we collected data about YouTube videos related to the COVID-19 pandemic and analyzed the patterns within each video's comments. Through the use of toxicity assessment, topic modeling, and social network analysis, we have detailed methods to 1)

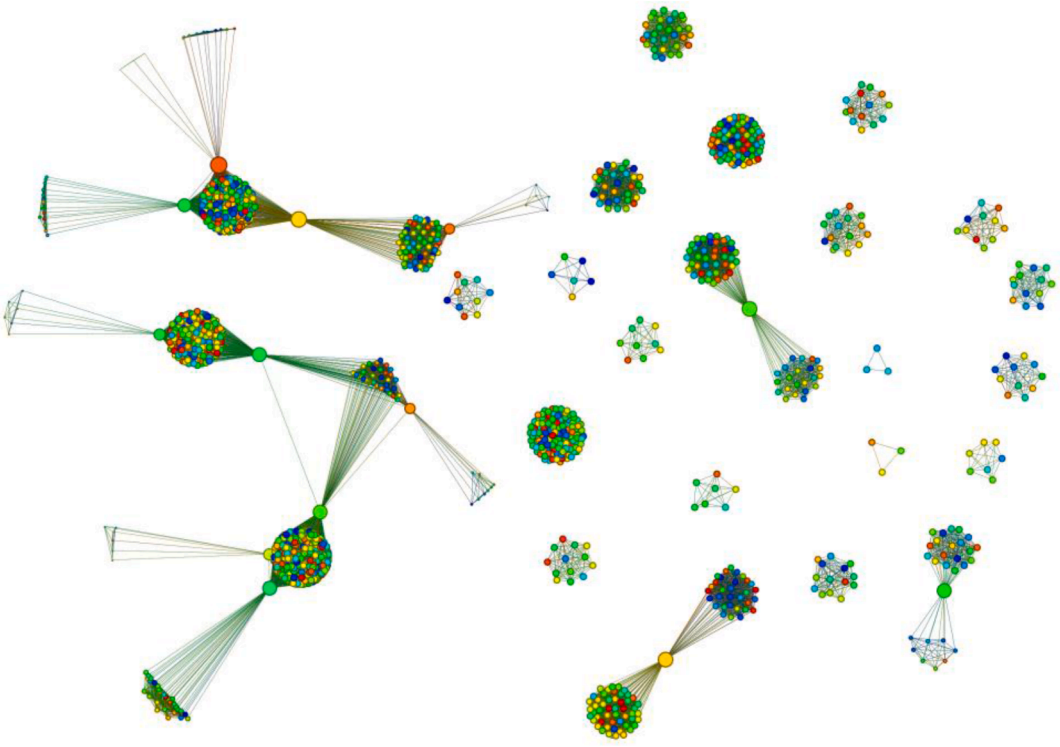

Fig. 19. Co-commenter shared video network with nodes (users) with toxicity score greater than 0.8 removed. Colors denote toxicity scores from lowest 0.5 (blue) to highest 1 (red).

Table 6
Summary of percentage in toxicity reduction in experimental simulation.

Experimental simulation	Toxicity score	Percentage reduction
Removal of top 10 users with high Betweenness centrality	0.720981759	0.21
Removal of top 10 users with high PageRank centrality	0.722317191	0.02
Removal of users with toxicity scores greater than 0.8	0.641927323	11.15

identify and measure toxicity of text content on OSNs, 2) identify the common topics of those toxic discussions, and 3) identify the commenters who are propagating that toxicity across a social network. Once we were able to understand the most toxic offenders and their behavior patterns, we were able to perform simulations to envision a network without them.

Methods such as the ones utilized in this work can be useful when incorporated into the moderation processes of OSNs. When toxic commenters and their behavior patterns are identified, administrators of those social media platforms can decide to either flag or remove such commenters from the network, which would improve the overall health of the communication platform by reducing its average toxicity. This technique can be applied to any social media platform. The issue remains, however, of that fine line between censorship and trampling the users' right to speech. It will be up to the administrators of these various OSNs to decide.

The methods used in this work can be generalized to other social media platforms in addition to YouTube without being inhibited by the varying content structure of those platforms. Toxicity assessment can be performed on any text corpus, whether it be YouTube video titles and descriptions, comments on YouTube videos, Tweets, Facebook posts, blog articles, comments on blog articles, Reddit threads, etc. LDA topic modeling can also be performed on any text corpus. Social network analysis can be performed on data from any social media platform as long as the researcher designates a "source" and a "target" upon input into the SNA tool. For example, for conducting SNA on Twitter data, a "source" input can be a unique user who tweeted a given tweet, and a "target" can be a set of users who retweeted that tweet.

Future work will include an expanded set of search terms to capture a wider range of discussions, and an extension of our analysis to conversations that have occurred since early May of 2020. Additional methods can be employed to discover patterns within the comments as well as among the behavior of each commenter. As the focus of this study was on YouTube activity, our work can be further validated by investigating similar behavior on other OSN platforms. We are also continually gathering new data and building on the dataset so that we can revise and expand our analysis over time. Social media guidelines restrict us from sharing our data, but we can make the initial data points available upon request so that future researchers and interested policymakers can collect the data from YouTube.

There are some potential limitations of this work. One type of vulnerability with machine learning algorithms is that an adversary can change the algorithm output by subtly perturbing the input, often unnoticeable by humans. Such inputs are called adversarial examples. We leveraged Perspective API, which is a state-of-the-art toxicity detection system developed by Google. However, there have been studies that showed that this tool can be deceived using adversarial inputs. Since then, Google has improved the API and we noticed a significant improvement in the API when we conducted this research. In part to improve its scoring, Jigsaw, Google and Wikipedia released their training and test data to international data competitors (Kaggle.com, 2018). Two goals were to set a new benchmark and to sample alternative solutions, including direct use of the Perspective API.

Author statement

AO and NA carried out the conception and study design.
AO carried out the data collection and analysis.
TK carried out the network analysis.
EM carried out the literature survey and helped with technical writing.
TM carried out the network analysis.
NA coordinated, helped draft the manuscript, and secured funding.
All authors contributed to writing and reviewing the manuscript. All authors approve the final manuscript.

Acknowledgements

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

8-Year-Old Boy Commits Suicide After Being Bullied | PEOPLE.com 2017 *8-Year-Old Boy Commits Suicide After Being Bullied | PEOPLE.com*. (2017). https://people.com/chica/8-year-old-boy-commits-suicide-after-being-bullied-at-school/.

Almerekhi, H., Kwak, H., Salminen, J., & Jansen, B. J. (2020). Are these comments triggering? Predicting triggers of toxicity in online discussions. In *Proceedings of the web conference 2020* (pp. 3033–3040). https://doi.org/10.1145/3366423.3380074. https://doi.org/.

Blei, D. M. (2003). *Latent dirichlet allocation*. 30.

Cacioppo, J. T., Fowler, J. H., & Christakis, N. A. (2009). Alone in the crowd: The structure and spread of loneliness in a large social network. *Journal of Personality and Social Psychology, 97*(6), 977–991. https://doi.org/10.1037/a0016076. PubMedhttps://doi.org/.

Chandrasekaran, R., Mehta, V., Valkunde, T., & Moustakas, E. (2020). Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study. *J Med Internet Res, 22*(10), e22624. https://doi.org/10.2196/22624. https://doi.org/.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (pp. 1217–1230). https://doi.org/10.1145/2998181.2998213. https://doi.org/.

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2015). Antisocial behavior in online discussion communities. In *, 9. Proceedings of the international aaai conference on web and social media*. https://ojs.aaai.org/index.php/ICWSM/article/view/14583.

Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine, 357*(4), 370–379. https://doi.org/10.1056/NEJMsa066082. https://doi.org/.

Christakis, N. A., & Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. *New England Journal of Medicine, 358*(21), 2249–2258. https://doi.org/10.1056/NEJMsa0706154. https://doi.org/.

Christakis, N. A., & Fowler, J. H. (2010). Social network sensors for early detection of contagious outbreaks. *PLOS ONE, 5*(9), e12948. https://doi.org/10.1371/journal.pone.0012948. https://doi.org/.

Christakis, N. A., & Fowler, J. H. (2013). Social contagion theory: Examining dynamic social networks and human behavior. *Statistics in Medicine, 32*(4), 556–577. https://doi.org/10.1002/sim.5408. https://doi.org/.

Coronavirus disease (COVID-19) – World Health Organization. (2019). WHO. https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media* (p. 11). https://ojs.aaai.org/index.php/ICWSM/article/view/14955.

Google (2011). https://web.archive.org/web/20111104131332/https://www.google.com/competition/howgooglesearchworks.html.

Fowler, J. H., & Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ, 337*, a2338. https://doi.org/10.1136/bmj.a2338. https://doi.org/.

Green, B., Horel, T., & Papachristos, A. V. (2017). Modeling contagion through social networks to explain and predict gunshot violence in Chicago, 2006 to 2014. *JAMA Internal Medicine, 177*(3), 326–333. https://doi.org/10.1001/jamainternmed.2016.8245. https://doi.org/.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is "Love": evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security* (pp. 2–12). https://doi.org/10.1145/3270101.3270103. https://doi.org/.

Han, X., & Tsvetkov, Y. (2020). Fortifying toxic speech detectors against veiled toxicity. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 7732–7739). https://doi.org/10.18653/v1/2020.emnlp-main.622. https://doi.org/.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current Directions in Psychological Science, 2*(3), 96–100. https://doi.org/10.1111/1467-8721.ep10770953. https://doi.org/.

Haven, A. P. (2013). *Florida cyberbullying: Two girls arrested after suicide of Rebecca Sedwick, 12*. October 15. The Guardian http://www.theguardian.com/world/2013/oct/15/florida-cyberbullying-rebecca-sedwick-two-girls-arrested.

Heidemann, J., Klier, M., & Probst, F. (2010). *Identifying key users in online social networks: A pagerank based approach*. 21.

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's perspective API built for detecting toxic comments. *ArXiv:1702.08138 [Cs]*. http://arxiv.org/abs/1702.08138.

Lee, S.-H., & Kim, H.-W. (2015). Why people post benevolent and malicious comments online. *Communications of the ACM, 58*(11), 74–79. https://doi.org/10.1145/2739042. https://doi.org/.

Märtens, M., Shen, S., Iosup, A., & Kuipers, F. (2015). Toxicity detection in multiplayer online games. In *2015 international workshop on network and systems support for games (NetGames)* (pp. 1–6). https://doi.org/10.1109/NetGames.2015.7382991. https://doi.org/.

Marcoux, T., Mead, E., & Agarwal, N. (2020). *The Ebb and flow of the COVID-19 misinformation themes*. 7.

Mednick, S. C., Christakis, N. A., & Fowler, J. H. (2010). The spread of sleep loss influences drug use in adolescent social networks. *PLOS ONE, 5*(3), e9775. https://doi.org/10.1371/journal.pone.0009775. https://doi.org/.

Mei, Q., Cai, D., Zhang, D., & Zhai, C. (2008). Topic modeling with network regularization. In *Proceedings of the 17th international conference on world wide web* (pp. 101–110). https://doi.org/10.1145/1367497.1367512. https://doi.org/.

Newman, M. E. J., & Girvan, M (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), Article 026113. https://doi.org/10.1103/PhysRevE.69.026113. https://doi.org/.

NW, 1615 L. St, Suite 800Washington, & Inquiries, D. 20036USA202-419-4300 | M.-857-8562 | F.-419-4372 | M. (2014, October 22). Online Harassment. *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2014/10/22/online-harassment/.

Obadimu, A., Mead, E., & Agarwal, N. (2019). *Identifying latent toxic features on YouTube using non-negative matrix factorization*.

Obadimu, A., Mead, E., Hussain, M. N., & Agarwal, N. (2019). Identifying toxicity within youtube video comment. In R. Thomson, H. Bisgin, C. Dancy, & A. Hyder (Eds.), *Social, cultural, and behavioral modeling* (pp. 214–223). Springer International Publishing.

Pavlopoulos, J., Thain, N., Dixon, L., & Androutsopoulos, I. (2019). ConvAI at SemEval-2019 Task 6: offensive language identification and categorization with perspective and BERT. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 571–576). https://doi.org/10.18653/v1/S19-2102. https://doi.org/.

Perez, C., & Germon, R. (2016). Chapter 7—graph creation and analysis for linking actors: application to social data. In R. Layton, & P. A. Watters (Eds.), *Automating open source intelligence* (pp. 103–129). Syngress. https://doi.org/10.1016/B978-0-12-802916-9.00007-5. https://doi.org/.

Qi, S., AlKulaib, L., & Broniatowski, D. A. (2018). Detecting and characterizing bot-like behavior on twitter. In R. Thomson, C. Dancy, A. Hyder, & H. Bisgin (Eds.), *Social, cultural, and behavioral modeling* (pp. 228–232). Springer International Publishing.

Rosenquist, J. N., Fowler, J. H., & Christakis, N. A. (2011). Social network determinants of depression. *Molecular Psychiatry, 16*(3), 273–281. https://doi.org/10.1038/mp.2010.13. https://doi.org/.

Lee, S. S., Chung, T., & McLeod, D. (2011). Dynamic item recommendation by topic modeling for social networks. In *2011 eighth international conference on information technology: New generations* (pp. 884–889). https://doi.org/10.1109/ITNG.2011.153. https://doi.org/.

Shachaf, P., & Hara, N. (2010). Beyond vandalism: Wikipedia trolls. *Journal of Information Science, 36*(3), 357–370. https://doi.org/10.1177/0165551510365390. https://doi.org/.

Sood, S. O., Antin, J., & Churchill, E. F. (2012). *Using crowdsourcing to improve profanity detection*. 6.

Srikanth, A. (2020). *Bullied 9-year-old Quaden Bayles passes on trip to Disneyland, donates money raised to anti-bullying charities instead*. February 28. TheHill [Text] https://thehill.com/changing-america/respect/diversity-inclusion/485137-nine-year-old-passing-on-trip-to-disneyland-and.

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior, 7*(3), 321–326. https://doi.org/10.1089/1094931041291295. https://doi.org/.

Varjas, K., Talley, J., Meyers, J., Parris, L., & Cutts, H. (2010). High school students' perceptions of motivations for cyberbullying: An exploratory study. *The Western Journal of Emergency Medicine, 11*(3), 269–273. PubMed.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media* (pp. 19–26). https://www.aclweb.org/anthology/W12-2103.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web* (pp. 1391–1399). https://doi.org/10.1145/3038912.3052591. https://doi.org/.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *2012 International conference on privacy, security, risk and trust and 2012 international confernece on social computing* (pp. 71–80). https://doi.org/10.1109/SocialCom-PASSAT.2012.55. https://doi.org/.

Yin, D., Xue, Z., & Hong, L. (2009). *Detection of harassment on web 2.0*. 7.