Journal of Computational Social Science (2020) 3:445–468
https://doi.org/10.1007/s42001-020-00087-4

RESEARCH ARTICLE

Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines

Joshua Uyheng[1] · **Kathleen M. Carley**[1]

Received: 23 July 2020 / Accepted: 3 October 2020 / Published online: 20 October 2020

Abstract

Online hate speech represents a serious problem exacerbated by the ongoing COVID-19 pandemic. Although often anchored in real-world social divisions, hate speech in cyberspace may also be fueled inorganically by inauthentic actors like social bots. This work presents and employs a methodological pipeline for assessing the links between hate speech and bot-driven activity through the lens of social cybersecurity. Using a combination of machine learning and network science tools, we empirically characterize Twitter conversations about the pandemic in the United States and the Philippines. Our integrated analysis reveals idiosyncratic relationships between bots and hate speech across datasets, highlighting different network dynamics of racially charged toxicity in the US and political conflicts in the Philippines. Most crucially, we discover that bot activity is linked to higher hate in both countries, especially in communities which are denser and more isolated from others. We discuss several insights for probing issues of online hate speech and coordinated disinformation, especially through a global approach to computational social science.

Keywords Hate speech · Social cybersecurity · Bots · Information maneuvers · COVID-19

Introduction

In the time of COVID-19, nations all over the world face not just a major public health crisis, but also a crisis of social relations [66, 82]. Especially in settings of entrenched inequalities and political polarization, the pandemic has exposed and exacerbated conflicts between social groups [31, 51]. In this work, we investigate how such dynamics play out in cyberspace. We specifically examine the

✉ Joshua Uyheng
juyheng@cs.cmu.edu

Kathleen M. Carley
kathleen.carley@cs.cmu.edu

[1] CASOS Center, Institute for Software Research, Carnegie Mellon University, Pittsburgh, USA

phenomenon of hate speech on social media, especially in relation to online disinformation [10, 20, 59]. In the context of a global pandemic, we ask to what extent the spread of online hate speech may be linked to bot-driven activities. We also probe what ends such information maneuvers may be instrumentalized toward, with consequences which extend beyond the digital sphere [5, 17, 70, 86].

This work bears several implications for understanding online hate speech in the context of the pandemic and beyond. We pivot from extant technical approaches of classifying hate speech [7, 35, 53], to theory-informed frameworks for characterizing it in the context of large-scale social interactions and potential information maneuvers [44, 80, 84]. We also reflect on the value of taking a global approach to computational social science, especially in the context of international issues like COVID-19, with its universal yet also distinct and unequal impacts in societies worldwide [31, 51, 54].

In the succeeding sections, we offer an overview of related work in this area to lay the conceptual foundations for this paper. First, we discuss the problem of hate speech on social media. We highlight the importance of shifting from the prevailing concern with classification to applied settings of characterization [80, 84]. Second, we link our view of hate speech to the literature on bots and information maneuvers. We specifically situate our work within the emerging field of social cybersecurity, which underscores the multidisciplinary nature of looking at disinformation in terms of sociotechnical systems [11, 15, 22]. Finally, we zero in on the present context of the global pandemic. We frame existing theoretical and methodological frameworks in terms of our real-world case studies of the United States and the Philippines, two countries which face rampant public health and social issues in the time of COVID-19 [26, 82].

Related work

Hate speech on social media: from classification to characterization

Hate speech has been broadly defined as abusive language that targets a specific group [30]. On social media, hate speech further proliferates in terms of highly connected 'highways', such that different forms and communities of hate intensify each other [21, 44]. Online hate speech, thus, goes beyond a merely linguistic phenomenon. It is also grounded in real-world divisions [64, 67]. Conversely, it also has real-world impacts. Empirical research shows that online hate can predict offline hate, especially against minorities like Black and Muslim populations [5, 86]. Probing the psychological processes underlying these relationships, scholars have also found that repeated exposure to hate speech may increase levels of prejudice by making individuals desensitized to derogatory and exclusionary narratives about certain groups [50, 70]. These processes consequently lead to broader strains on intergroup relations and heightened political polarization [17].

Identifying hateful online behaviors, thus, represents an important task toward safeguarding the health of digital platforms and supporting community-based responses to countering toxicity [38, 55]. Major efforts from a computational and

natural language processing standpoint have focused on the task of automatically classifying hate speech [25, 35, 85]. Extensive progress has been made with regard to harnessing cutting-edge deep neural models to predict hate speech in text with a high degree of accuracy [7]. For example, innovative recent work trains word embedding approaches to account for the specific language of 'othering' used in hate speech to further improve prediction performance [2, 32].

Emerging literature in the social and computational sciences increasingly adopts the view that interpretability of models is likewise important [53]. In certain cases, the complex features of state-of-the-art neural models may make it challenging for researchers to understand the predictions made. From this standpoint, valuable approaches have also looked at simpler models which rely on theoretically enriched and engineered features. For instance, a valuable research area in psycholinguistics documents systematic and cross-cultural links between various patterns of word usage with behaviors and mental states [63, 76]. By employing more interpretable features, it becomes more straightforward to describe the social and communicative dynamics which underpin hate speech as it is used in context. This enables researchers to better understand not just which texts may be linked to hate speech, but also how they communicate hate, evolve in communities, and reinforce conflicts [3, 33, 47, 50].

Hate speech in bot-driven information maneuvers

Building on this broader research agenda of hate speech characterization, we further situate its spread within the context of potential information operations. Although digital platforms have often been hailed for their capacity to democratize public communication, they have also been scrutinized for the ways they facilitate disinformation by inauthentic actors such as social bots [10, 59]. Especially in recent years, social bots have played a noteworthy role in proliferating digital content pollution across a variety of contentious settings [34, 38, 69].

Given their ubiquity, social bots have also been defined in heterogeneous ways. A major definition in this area refers to bots as algorithms designed to automatically generate content and interact with human users [34]. Scholars note that many kinds of social bots display distinctive behavioral patterns to achieve concerted political or economic ends, including spamming, increasing the perceived following of a politician, or sowing discord [83, 87]. The literature also points out that some bots—sometimes called cyborgs—are only partially automated, such that certain messages they send can attain a higher level of sophistication due to intermittent interventions by humans during real-time interactions [27, 80]. For the purposes of this work, and in line with the operationalizations embedded in tools described in Sect. 3.3, we align with this broadest sense of a social bot. Hence, social bots in this study refer to accounts using at least some level of automation to achieve informational objectives in the context of online social networks.

Social cybersecurity presents a computational social science framework for organizing the activities of social bots around specific information maneuvers [22]. Information maneuvers are broadly divided into two categories: narrative and

network maneuvers [11, 14, 15]. These are further subdivided into specific positive and negative strategies under the *BEND* framework as a heuristic. Narrative maneuvers have to do with shaping the way certain storylines play out in public discourse. Coordinated actors may seek to positively reinforce or negatively shut out certain perspectives in the online conversation [61, 75]. Positive maneuvers include the engagement of related storylines, the further explanation of a topic, or the excitement and enhancement of a group (E). Negative maneuvers, on the other hand, seek to dismiss storylines, distort messages, or dismay and distract the concerned group (D).

Network maneuvers, on the other hand, aim to influence the structural features of social media interactions. By altering who talks to whom, information maneuvers may strategically shape the flow of information regardless of its content [8, 37, 57]. Bots targeting such objectives may seek to make certain actors more influential or less influential. They may also aim to break up a community or inorganically link two communities together [4, 83]. These notions correspond to the positive maneuvers summarized as backing, boosting, bridging, and building (B); and negative maneuvers given by neutralizing, nuking, narrowing, and neglecting (N). Such maneuvers are instrumentalized toward diverse ends, including manipulating public opinion and exacerbating political unrest [16, 43, 78, 78].

This paper, thus, advances a view of hate speech as a complex social phenomenon which may be embedded in information maneuvers. Without discounting that the broader public may organically engage in hateful talk online [71, 72], the evidence that bots exert significant influence in driving digital toxicity nonetheless makes it important for researchers to examine their potential impacts in the context of the pandemic [74, 79]. For instance, from a BEND perspective, bots can promote networked hate by building hate groups and backing hate-promoting opinion leaders. In addition, they could shape narratives through various maneuvers such as distort or dismay to increase the volume of messages spreading hate. We therefore conceptualize bot-driven hate in line with an analytical perspective harnessing both the computational and social sciences [49, 52, 82].

Global hate during the pandemic: The United States and the Philippines

Against the backdrop of the COVID-19 pandemic, the foregoing insights become important on a global scale. Emerging studies have looked at how the pandemic has fueled racialized and xenophobic tensions [31, 51]. In particular, because the first notable outbreaks happened in China, it has become a major concern whether online talk stokes sinophobia [73, 89]. Months into the pandemic, the objects of conflict have also evolved, accounting for the politically heated landscape of institutional pandemic response and growing disparities in equitable treatment delivery [42, 54].

Within specific geopolitical landscapes, the dynamics of online hate and disinformation may vary. In the United States, robust scholarship tackles the potential influence of foreign influence campaigns [6, 9]. Known operations, for example, have stoked online conflicts in relation to wider racial divisions [4, 89]. Meanwhile, in the Philippines, international tensions with China may also play a role in shaping

racially charged online hate [58, 78]. However, in view of well-documented evidence that disinformation in the Philippines is primarily domestic and state-sponsored, information maneuvers may also play a role in primarily political conflicts [62, 79].

Common between the two countries, however, are a combination of hyperpartisanship among polarized publics, the rise in populist-authoritarian leadership, and sustained challenges in curbing the pandemic [1, 68, 81]. At the time of writing in mid-September, the US faces the largest number of COVID-19 cases in the world, while the Philippines has some of the highest in Southeast Asia [88]. Taken together, these factors may contribute to the proliferation of hate speech during a volatile political and economic historical moment, especially towards locally marginalized groups.

In this work, we seek not only to characterize hate speech in the context of potential bot-driven activities, but also to compare their dynamics across two distinct geopolitical settings. This affords us a more holistic and diverse view of the digital landscape amid a planetary crisis like COVID-19. It also helps us advance a more complete practice of computational social science in extending beyond the predominant focus on Western, Educated, Industrialized, Rich, and Democratic nations (WEIRD) [39, 41]. These bear implications we discuss further in our concluding sections.

Data and method

Data collection

Online conversations around the COVID-19 pandemic were collected using Twitter's REST application programming interface (API) updated on a daily basis. We used search terms about the pandemic in the Philippines and the US. Localized hashtags were used to delineate tweets of interest, specifically '#COVID19PH' for the Philippines and '#COVID19US' for the US. To enforce reasonable comparability between the datasets, hashtags which were not geographically specific to either the US or the Philippines were not included, such as #Wuhanvirus, #Chinavirus, and #coronavirus, which related studies have otherwise used [36, 89]. In this manner, we sought to delineate our study to discourse surrounding the mainstream, country-specific hashtag related to the topic. For the comparative purposes of our study, all data on the US and all data on the Philippines were processed in parallel; that is, datasets were treated separately from each other with comparisons drawn after analytical strategies were implemented.

We continued data collection over a 75-day period from March 5 to May 19 of 2020. For parsimony, data were strategically time-bound before the emergence of the '#BlackLivesMatter' protests, which also impacted public discourse worldwide. We stored user metadata, tweet metadata, data on user interactions, and data about the hashtags and URLs each tweet used. Twitter interactions in this work collectively refer to the sum of retweets, replies, mentions, and quotes. Retweets were included

to capture the intuition that they account for the amplification of certain messages over others. Hashtags were also retained for all subsequent textual analysis.

The final datasets consisted of 12 million tweets featuring 1.6 million users in the US, and 15 million tweets featuring 1 million unique users in the Philippines. Larger datasets have been collected with more universal and dynamically updated search terms [24]. However, for the purposes of this work, it was important to hold constant the geographically specific nature of the conversation; thus, we maintained this smaller, yet more theoretically appropriate dataset for analysis [60]. It is important to note that state-level terms were not included in the US collection stream; thus, the data on the US may represent a smaller fraction of American pandemic talk, focusing on people discussing the pandemic in relation to the nation at large. Future work may probe whether the bot-driven hate dynamics found here extend more broadly.

Hate speech classification

Hate speech scores were assigned to each tweet using a machine learning algorithm. For user-level analysis, we obtained the average hate speech score assigned to all tweets by the account within the time interval under consideration. Our model utilized handcrafted linguistic features using the Netmapper software [11, 79, 80]. These features included lexical counts of pronouns, abusive terms, exclusive terms, absolutist terms, and identity terms, among others based on prior scholarship linking language use to psychological states [63, 76]. Second-order interactions were also added to capture the co-occurrence of various linguistic features in a given tweet. We performed this feature enrichment in a multilingual setting, capturing counts for English words as well as common languages in the Philippines (e.g., Tagalog, Cebuano).

Our model was trained on a seminal benchmark dataset for hate speech [30]. This dataset had three labeled classes for training in a supervised setting: (a) hate speech, (b) offensive speech, and (c) regular speech. A crucial consideration here is the distinction between hate speech and offensive speech, as the latter may include abusive terms which are not necessarily targeted toward any specific groups. By generating predictions for both labels, we mitigate false positives. However, we note that this may also conversely impose overly stringent predictions for hate speech, resulting in some level of underprediction [30]. For this purpose, we also consider predictions of offensive language in some of our analysis.

Across several experiments, the best model was a random forest classifier achieving over 80% performance in terms of micro-averaged and weighted F1 to account for class imbalance. For the same reason, training was also performed with oversampling on the training set. While more performant models could be designed using state-of-the-art neural network architectures, we followed practical arguments in recent work advocating interpretability to perform hate speech prediction at scale with theoretically motivated features [53].

Table 1 summarizes the performance of our chosen model relative to a random baseline, a heuristic baseline, and a regularized logistic regression baseline. The random baseline was evaluated as the average performance obtained using a random

Table 1 Evaluation of simple classifiers for hate speech using second-order psycholinguistic features

Classifier	Micro F1	Weighted F1
Random baseline	0.6262	0.6269
Heuristic baseline	0.6880	0.7236
Logistic regression	0.6342	0.6961
Random forest	**0.8417**	**0.8293**

Numbers in bold indicate the model with the best predictive performance

Subsequent analysis uses the best-performing random forest model

permutation of true labels. Anchored in the conceptual definition of hate speech [30], the heuristic baseline is set up as follows: we predict hate speech if there is at least one abusive term and one identity term, we predict offensive language if there is at least one abusive term but no identity terms, and we predict regular speech otherwise. For logistic regression, we tested different values for the regularization parameter. For the random forest model, we tested different numbers of estimators from 5 to 100 in increments of five.

Bot prediction

Bot scores were also assigned to each user using the BotHunter algorithm [11, 13]. BotHunter also relies on a random forest classifier trained on several labeled datasets of known social bots. Adopting a tiered structure, BotHunter progressively leverages account information, network information, and dynamic features to generate performant bot predictions. It has been successfully used to understand various domestic and international influence campaigns [14, 16, 78, 80].

BotHunter was the preferred bot prediction algorithm in this study due to its scalability and reliable accuracy over large and diverse datasets. In the context of this study, we do not retrain the BotHunter model. However, BotHunter uses a variety of annotated examples from diverse domains, including a bot attack on NATO, known Russian accounts associated with the Internet Research Agency (IRA) which interfered in the 2016 US elections, and a large dataset of known suspended accounts [12]. Experiments show good domain generalizability across a variety of different types of bots in different contexts. Detailed training specifications and generalized performance on different domains are examined in prior work, but are beyond the scope of the present study [11, 12].

Identity analysis

To determine the targets of hate speech, we employed an identity-based approach [17, 50]. Past research had developed a comprehensive lexicon of identity terms which have also been made available on the Netmapper software [23, 45]. Each identity term in the lexicon was further subdivided into four classes: political identities (e.g., senator, president), gendered identities (e.g., women, transgender), racial/nationality identities (e.g., Black, Filipino), and religious identities (e.g., priest,

imam). Subclasses were based on related work on hate speech and social media identity prediction more generally [46, 65]. Acknowledging that identities are intersectional [28], this subcategorization scheme was set up such that each identity term could be assigned any combination of the four identities mentioned above. Each tweet was also assigned an identity score based on the number of times it invoked each subclass of identity term.

Social network analysis

We use the ORA software to conduct social network analysis on our datasets [23]. For a given time t, we represented the Twitter conversation for a given country c using a graph $G_t^c = \left(V_t^c, E_t^c\right)$. For the purposes of the present analysis, we divided our data into daily datasets. Here, V_t^c represents the set of Twitter accounts represented in the data at the specified time, which act as nodes in the network. Meanwhile, E_t^c represents the edges between nodes, representing the interactions between users. Edges were weighted by the sum of all Twitter interactions within the time period consisting of quotes, mentions, replies, and retweets.

Network influence

We used the network representations described above to obtain measures of degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality using the ORA software [11, 19, 40]. Centrality metrics on social networks capture varied notions of how important each actor is in the spread of information.

Total degree centrality in this context captures the extent to which users are engaged in all kinds of Twitter interactions. Betweenness centrality denotes how much accounts act as bridges, thereby connecting relatively distinct groups. Eigenvector centrality corresponds to the degree to which certain accounts interact with other influential accounts. Finally, closeness centrality gives us an idea about how much agents control overall information flow in the network. Results pertaining to network influence paint a holistic picture of whether more hateful users were successful in accruing network influence, especially if they were more bot-like.

Cluster analysis

Finally, to view hate speech and bot activity in the context of localized interactions, we performed community detection using the Leiden algorithm [77]. The Leiden algorithm advances the state of the art in network clustering over the more commonly used Louvain method [18]. Intuitively, it aims to partition a given social network into smaller groups such that agents in the same cluster are more likely to interact with each other than with agents from other clusters. With guarantees to produce well-defined clusters, the Leiden algorithm automatically selects the optimal number of communities in a network. It also boasts faster runtime even with large datasets.

Localizing our analysis in terms of Leiden groups conceptually aligns with broader understandings of social media dynamics. Despite the large-scale nature of online conversations, most agents are exposed only to information within specific communities [8, 37, 79]. Thus, it is valuable to examine how bot activities and hate speech co-vary on the cluster level alongside the global analysis offered by the centrality measures discussed above.

Clusters may additionally be characterized by a variety of metrics. Size is a basic measurement of how many agents have been assigned to a given cluster. The E/I index provides a measure of how exclusively agents in a cluster interact with each other versus those in other clusters [48]. Values closer to +1 indicate that the cluster interacts with agents outside their cluster; whereas values closer to -1 suggest that agents are siloed in their interactions. Finally, the Cheeger constant measures the amount of bottleneck behavior, with higher values indicating that information may be concentrated among a smaller number of nodes relative to the rest [56]. We implement several regression models to determine the relationship of these community features with bot activity and hate speech.

Results

Overview of bot and hate scores

Our first set of results presents BotHunter and hate speech predictions on our two datasets. Table 2 suggests that levels of bot activity and hate speech are comparable across the two contexts. Using a standard 80% BotHunter threshold [78, 79], we note that while a larger proportion of tweets are produced by bots in the US dataset, we captured more bot tweets in raw numbers in the Philippine dataset. Conversely, while a larger number of bot users were found in raw numbers in the US, the Philippines had a greater proportion of bot users. Collectively, these initial predictions suggest that bots were tweeting at a more prolific rate in the Philippines compared to the US.

Table 2 Summary of datasets with BotHunter, hate speech, and offensive speech scores

Dataset	Bot predictions		Hate scores		Offensive scores	
	Tweets	Users	Tweets	Users	Tweets	Users
US	3.026 M (26.31%)	237 K (14.91%)	0.1023 (0.0769)	0.1073 (0.0656)	0.2773 (0.1519)	0.2914 (0.1418)
PH	3.436 M (21.73%)	150 K (15.70%)	0.0896 (0.0717)	0.0924 (0.0600)	0.2672 (0.1375)	0.2836 (0.1180)

For this table, an 80% probability threshold was used to classify a user as a bot. Bot tweets refer to tweets produced by bots predicted by the same threshold. Parentheses provide dataset proportions for BotHunter predictions and standard deviations for hate speech and offensive speech scores

Table 2 shows the mean scores obtained for both hate speech and offensive speech. Average scores were consistently higher in the US for hate speech and offensive speech, both on the tweet level and on the user level. Overall, however, hate speech and offensive speech distributions are skewed to the right, although the average value for offensive speech is higher than that of hate speech. This indicates that, despite the relatively neutral choice of search terms used during data collection, offensive talk is relatively common in both datasets. Cases of hate speech with high probability, however, are quite rare [89].

Four months after data collection, we found that a notable percentage of accounts were no longer active on Twitter. Suspensions accounted for 5.96% of the accounts in our US dataset, and 5.18% of the accounts in our Philippine dataset. Offering some validation for our model predictions, Welch two-sample t tests established that in the US, suspended accounts had higher BotHunter probabilities ($t = 41.82, p < 0.001$), and higher scores on offensive speech ($t = 8.04, p < 0.001$). Interestingly, no significant difference was found in levels of hate speech between suspended and non-suspended accounts in the US. Meanwhile, in the Philippines, we saw that suspended accounts consistently had higher BotHunter probabilities ($t = 18.78, p < 0.001$), hate speech scores ($t = 6.55, p < 0.001$), and offensive speech scores ($t = 9.43, p < 0.001$).

Taken together, we observe different patterns between the two datasets in predicting whether an account was suspended based on these factors. Figure 1 visualizes the coefficients of a logistic regression model which predicts the binary suspension outcome based on BotHunter probabilities, hate speech scores, and offensive speech scores. Here, we see that accounts in the US dataset were more likely to be suspended if they had higher bot probabilities. Higher hate speech scores and higher offensive speech scores also predicted suspension in the US, but not as strongly for accounts that were also bot-like. Meanwhile, in the Philippines, higher suspension probability was observed for accounts that were bot-like, that expressed offensive speech, or that were both bot-like and expressing hate speech.

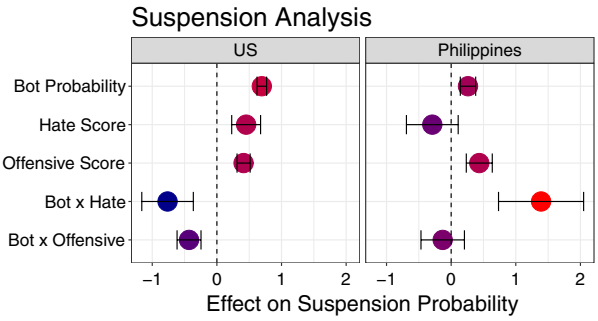

Fig. 1 Coefficients of a logistic regression model predicting whether an account is suspended based on its BotHunter probability, its hate speech score, and its offensive speech score. Error bars represent 95% confidence intervals

Table 3 Tweet- and user-level correlation coefficients between BotHunter scores and hate speech/offensive speech scores

Dataset	Bot-to-hate		Bot-to-offensive	
	Tweets	Users	Tweets	Users
US	− 0.0365***	− 0.0161***	− 0.0518***	− 0.0124***
PH	− 0.0030***	− 0.0238***	− 0.0242***	− 0.0066***

Tweet-level and user-level correlations

On both the basic tweet and user levels, we find no evidence of correlation between bot activity and levels of both hate speech and offensive speech. Pearson correlation coefficients estimated for all relationships are summarized in Table 3, all indicating near-zero values with statistical significance. Thus, overall, tweets from more bot-like users were not necessarily more hateful or more offensive.

On balance, these initial results suggest that on a basic tweet or user level, clear conclusions are difficult to draw about the behavior of bots in relation to hate speech. In both the US and the Philippines, the large-scale public discourse triggered by the pandemic encompasses massive numbers of participants. Considering our findings, we observe that moments of crisis may trigger relatively common offensive or inappropriate content, but tweets by bots will not be the only ones expressing toxicity. Hence, we suggest the importance of accounting for the localized interactional settings in which both bot activity and hate speech occur. We tackle these questions in the succeeding sections.

Are hateful bots more influential?

In the succeeding prose, we use the term high-hate bots to refer to users with BotHunter scores above our designated threshold of 80%, and hate scores in the highest quartile in our dataset. We assume that users with BotHunter scores with less than 80% are humans. Furthermore, we consider low-hate users based on the lowest quartile of hate speech scores; and mid-hate users in terms of the remaining two quartiles of the dataset, respectively. Through this approach, we identify low-hate bots and mid-hate bots, as well as humans with low-hate, mid-hate, and high-hate designations. We then examine their mean centrality scores to assess their relative influence in the dataset.

Figure 2 summarizes these results. For comparability, values are expressed relative to the maximum for each centrality score, which is linearly normalized to 1. Highlighting our social cybersecurity lens, we look at how different patterns of betweenness, closeness, eigenvector, and total degree centrality correspond to different relationships of bot activity and hate speech.

Bot hate and influence in the US

In the US, we find that low-hate bots have higher betweenness centrality than low-hate humans, but high-hate bots have lower betweenness centrality than low-hate

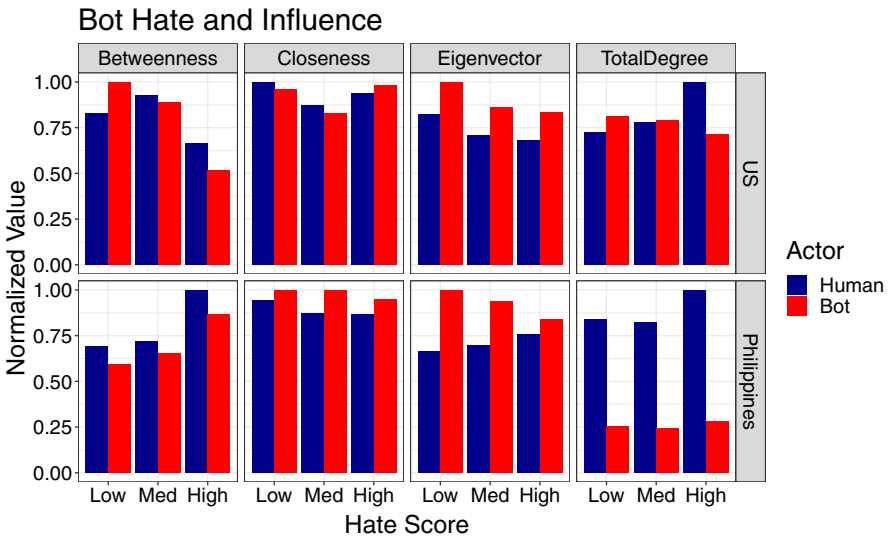

Fig. 2 Normalized mean centrality of users in the datasets. Users are divided into low, medium, and high hate scores, respectively defined as the bottom quartile, the middle 50%, and the top quartile of hate scores. Bot predictions are based on an 80% probability threshold

bots. Hence, it appears that bots act more successfully as bridges when they express low levels of hate, but when they do express high levels of hate, they are more embedded within a single local neighborhood of the social network.

Closeness centrality suggests that the most well-distributed actors in the network—and, therefore, holding significant influence over network information flow—are low-hate humans and high-hate bots. This suggests two major faces of the US conversation about the pandemic, with one view controlled by organic, non-hateful users, and the other controlled by inorganic, hateful users.

Bots also consistently appear to have higher eigenvector centrality than their human counterparts, regardless of their level of hate. Thus, it seems that bots in the US conversation are interacting with other accounts which interact actively with others. These behaviors broadly align with positive network maneuvers under the BEND framework, whereby bots may seek to amplify their messages by swarming the interactions of network influencers.

Finally, we see that as hate increases, the total degree centrality of bots steadily decreases. Conversely, especially for high-hate humans, the total degree centrality of humans increases markedly. Hence, the most active accounts in propagating hate within Twitter interactions in the US are not actually bots, but humans.

Bot hate and influence in the Philippines

Meanwhile, in the Philippines, centrality measures feature consistently different patterns, suggesting a distinct information landscape of hate and bot activity. Overall, humans had higher betweenness centrality and total degree centrality than bots,

regardless of their level of hate. The former result means that humans acted more as bridges between groups in the Philippine conversation, while bots were more likely to be focused on a single local neighborhood of the social network. The latter result, on the other hand, indicates that humans were also consistently the most engaged in all kinds of interactions with others.

That being said, we note that both betweenness centrality and total degree centrality also steadily increased as levels of hate increased. Hence, regardless of whether the account was a bot or a human, more hateful users served more powerful bridging functions between groups, and engaged in higher overall levels of Twitter interaction with others.

For closeness centrality, we observed only minor differences across the categories under observation. In general, however, bots tended to have higher closeness centrality than their human counterparts at comparable levels of hate. This indicates that information flow in the Philippine Twitter conversation around the pandemic tended to be controlled slightly more by inauthentic accounts, as these were more well-distributed throughout the social network.

A similar picture is seen in terms of eigenvector centrality. Bots also consistently had higher eigenvector centrality than their comparable human counterparts, indicating that they interacted with other influential others more often. But as hate increased, bots had lower eigenvector centrality; humans, on the other hand, had higher eigenvector centrality. This suggests that more hateful bots were less likely to interact with influencers in the conversation, but more hateful humans were more likely to do the same.

Whom do hateful bots target?

In our third set of results, we further characterize bot–hate dynamics during the pandemic in terms of the identity groups that they target. Following the general understanding of hate speech as toxicity aimed toward specific groups, higher mentions of specific identity types among more bot-like and hateful users are suggestive of more consistent targeting. Our measurements are summarized in Fig. 3. We use the same nomenclature to describe low-hate, mid-hate, and high-hate bots and humans. No normalization is conducted here, however, since all measures are on the same scale and order of magnitude. This additionally facilitates ordinal comparison of the most targeted identity groups.

As expected, for both datasets, more hateful users mention more identity terms across all categories. Based on a series of one-way ANOVAs, the results are consistent across gendered (US: $F = 13940, p < 0.001$; Philippines: $F = 103235, p < 0.001$), religious (US: $F = 2883, p < 0.001$; Philippines: $F = 74937, p < 0.001$), political (US: $F = 96901, p < .001$; Philippines: $F = 168914, t < 0.001$), and racial identities (US: $F = 368997, p < 0.001$; Philippines: $F = 56364, p < 0.001$). Generally, bots also follow the same ordinal trend with regard to which identity subclasses are targeted for both the US ($r = 0.996, p < 0.001$) and the Philippines ($r = 0.999, p < 0.001$). In other words, within the same country, bots and humans are discussing the same identity targets at similar levels.

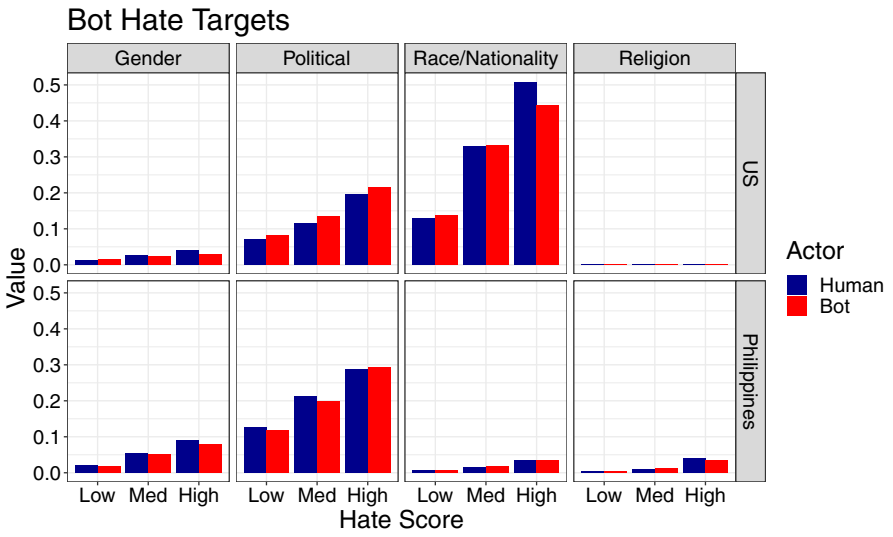

Fig. 3 Mean identity score of users in dataset. Users are divided into low, medium, and high hate scores, respectively defined as the bottom quartile, the middle 50%, and the top quartile of hate scores. Bot predictions are based on an 80% probability threshold

In the US, the most salient identity category has to do with race and nationality. Less bot-like, yet high-hate accounts mention race and nationality identities more than their more bot-like counterparts. But among low-hate users, bots mention race and nationality more than humans. This may indicate that more hateful users are already engaged in racially charged discussion, while inauthentic accounts may seek to boost such discussions among non-hateful users.

Bots in the US are also more likely to mention political identities across the board, especially in more hateful messages. This points to more concerted attempts by inauthentic accounts to heighten politically charged conflicts amid the pandemic. For gendered and religious identities, we observe relatively low values for all categories, indicating that these may not be salient targets of conversation in either hateful or less hateful talk.

In the Philippines, on the other hand, political identities comprise the most prominently discussed group. Political identities featured especially in more hateful discussions, especially when users were more bot-like. Even among non-hateful users, however, political identities were likewise frequent targets of discussion, even more among humans than among bots. None of the other identity categories had particularly prominent discussion across more and less bot-like and hateful users. It is particularly noteworthy here that the salience of race and nationality is quite low, in comparison to its ubiquity in the US discussion.

Extending this comparative standpoint, it is also worth remarking that political identities—the most salient category in the Philippines—is discussed with identity scores in the 0.20–0.30 range for mid-hate and high-hate users. In contrast, identities associated with race and nationality are discussed in the US at a much higher rate with scores over 0.40 among high-hate users. This indicates that the difference in the

targeting behavior of hate speech between countries is not just qualitative. It is also quantitative, with topics of race and nationality brought up over 50% more in the US than the most contentious political identities in the Philippines.

Does bot activity predict community hate?

We now consider whether there exist links between bot activity and hate speech on a community level. As shown in Sect. 4.2, no strong relationship exists on a tweet level or a user level for bot activity and hate speech. Literature on the community-level dynamics of social media interactions, however, suggests that attention to localized relationships may unearth valuable insights for social cybersecurity [15, 79].

Thus, we explore the relationship between hate speech and bot activity on the level of Leiden groups. Trivial clusters made up of isolates and pendants were removed to avoid analysis of degenerate groups. Given each social network for a given day, we note that Leiden grouping produced an average of 898 clusters in the US, with overall numbers ranging from 105 at the minimum to 2620 at the maximum. In the Philippines, Leiden grouping produced an average of 347.5 clusters, ranging overall from 44 to 1485 clusters on any given day in our dataset.

Accounting further for network maneuvers under the BEND framework, we set several cluster features described in Sect. 3.5 as additional predictors. We conduct this analysis over all Leiden groups derived for each time period, without accounting for temporal dynamics. These may be explored in subsequent work.

Table 4 summarizes the findings from our regression analysis. Community-level scores for hate and bot activity are computed as follows. We first let $h_{i,1}, h_{i,2}, \ldots, h_{i,n}$ be the predicted hate scores for each user in Leiden group i with n_i accounts. Next, let $b_{i,1}, b_{i,2}, \ldots, b_{i,n}$ likewise represent their BotHunter scores. Then, the community-level hate score is given by $\bar{h_i} = \frac{1}{n} \sum_j h_{i,j}$ and the community-level bot score is given by $\bar{b_i} = \frac{1}{n} \sum_j b_{i,j}$. Model 1 reports the base model without interaction effects. Model 2 includes all interactions between cluster features and community-level bot score. Model 3 optimizes the number of coefficients for parsimony by stepwise AIC backward selection. Results on both countries are reported separately.

Across both the US and the Philippines, we uncover a distinct finding: that higher community-level hate is predicted by the interaction of high community-level bot activity and high cluster density. These effects are significant in Model 2 for each country (US: $b = 0.065, \text{SE} = 0.019, p < 0.001$; Philippines: $= 0.169, \text{SE} = 0.031, p < 0.001$), as well as in Model 3 after AIC backward selection (US: $b = 0.065, \text{SE} = 0.019, p < .001$; Philippines: $b = 0.166, \text{SE} = 0.031, p < 0.001$). Overall, this suggests that bots spread hate most effectively in tightly connected groups.

Optimized in Model 3, the commonality of this finding across the two countries is especially important given the main effects of bot score (US: $b = -0.103, \text{SE} = 0.021, p < .001$; Philippines: $b = -0.041, \text{SE} = 0.007, p < 0.001$) and density (US: $b = -0.045, \text{SE} = 0.011, p < 0.001$; Philippines:

Table 4 Summary of regression analysis for community-level bot activity and hate

Variable	Model 1	Model 2	Model 3
US			
Intercept	0.096 (0.007)***	0.136 (0.038)***	0.143 (0.013)***
Bot Score	− 0.016 (0.001)***	− 0.092 (0.063)	− 0.103 (0.021)***
Size	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Density	− 0.009 (0.003)**	− 0.045 (0.011)***	− 0.045 (0.011)***
E/I Index	− 0.034 (0.007)***	− 0.002 (0.038)	0.005 (0.013)
Cheeger Score	0.011 (0.008)	0.025 (0.056)	0.016 (0.008)
Interactions			
Bot × Size	–	0.000 (0.000)	–
Bot × Density	–	0.065 (0.019)***	0.065 (0.019)***
Bot × E/I	–	− 0.063 (0.063)	− 0.074 (0.022)***
Bot × Cheeger	–	− 0.017 (0.092)	–
Philippines			
Intercept	0.068 (0.007)***	0.063 (0.059)	0.087 (0.008)***
Bot Score	− 0.017 (0.002)***	− 0.010 (0.098)	− 0.051 (0.007)***
Size	0.000 (0.000)	0.000 (0.000)	–
Density	0.003 (0.005)	− 0.093 (0.018)***	− 0.091 (0.018)***
E/I Index	− 0.041 (0.007)***	−0.066 (0.059)	− 0.042 (0.008)***
Cheeger Score	0.028 (0.007)***	0.094 (0.069)	0.064 (0.025)*
Interactions			
Bot × Size	–	0.000 (0.000)	–
Bot × Density	–	0.169 (0.031)***	0.166 (0.031)***
Bot × E/I	–	0.031 (0.099)	–
Bot × Cheeger	-	− 0.108 (0.114)	− 0.058 (0.038)

Coefficients reported are unstandardized.

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

$b = -0.091, \text{SE} = 0.018, p < 0.001$) alone. These estimates suggest that higher levels of bot activity in a non-dense community may not be associated with high hate scores; conversely, highly dense communities without much bot activity may also not be associated with more hate speech. Instead, it is the combined presence of high bot activity and frequent internal interactions predict greater levels of hate.

Furthermore, in the US, Model 3 shows that higher community-level hate is likewise predicted by the interplay of increased bot activity and decreased E/I indices ($b = -0.074, \text{SE} = 0.022, p < 0.001$). This further underscores the community-level dynamics of bot-driven hate. Here, higher bot activity in clusters with low E/I indices suggests salient bot interactions in echo chamber-like groups. Bots in the US conversation thus spread hate in dense and isolated groups. We additionally note that these effects are derived controlling for cluster size and Cheeger scores. These respectively indicate that the observed effects were consistent across differently sized clusters; and that whether or not interactions were concentrated on a few

influential accounts did not significantly dictate the relationship between bots and hate speech.

Finally, in the Philippines, Model 3 shows a main effect for E/I index ($b = -0.042, \text{SE} = 0.008, p < 0.001$). This indicates that hate proliferates in the Philippines in echo chamber-like communities. This pattern is moreover irrespective of bot activity given that the backward selection process removes the interaction effect between bot scores and the E/I index entirely. A significant main effect was also detected for Cheeger scores ($b = 0.064, \text{SE} = 0.025, p < 0.05$). Thus, bottlenecking behavior also predicts more hate; that is, in communities where discussion is dominated by a few users, hate is more likely to proliferate.

Discussion

This study presented an empirical analysis of online hate speech in relation to bot activity in the context of the COVID-19 pandemic. Comparing case studies of the US and the Philippines, we found that bot-driven hate features idiosyncratic dynamics across contexts, but may also be linked to common sociotechnical processes. In the succeeding sections, we discuss the implications of these findings in relation to the broader literature, highlighting our contributions not only to extant scholarship on hate speech and disinformation, but also to computational social science more broadly.

Integrative approaches to online hate and disinformation

Prevailing approaches to online hate speech have focused on identifying it as a linguistic phenomenon [7, 35, 84, 85]. In this work, we show how these methods can be reoriented toward the more situated task of characterizing hate speech in the concrete setting of understanding COVID-19 discourse [53]. Distinct studies have previously focused on the targeted nature of hate speech [2, 32, 33], its spread in communities [44, 47], and the potential role of social bots [69, 74]. Our work demonstrates a unified framework for viewing these phenomena as interlinked processes, thereby generating rich insights by examining their interplay.

In comparison to related studies, we found that our estimate of bot activity represented slightly lower levels than the proportions found by a related study on the COVID-19 pandemic, which predicted that 40.4% of all messages were bot-generated [36]. In a more extreme case, another study linked to spam activity surrounding financial microblogs posited an even higher proportion of around 71% bot activity [29]. However, our estimates do remain higher than the baseline reported in an earlier landmark study of bot activity, which estimates bot activity at around 9–15% [83]. Hence, our findings concur with the overall assessment that bot activity is higher during the pandemic relative to a general baseline. We surmise that the lower proportion may be due to the mainstream hashtags used in our data collection, whereas adjacent work on the pandemic has consciously

analyzed hashtags with potentially incendiary implications to begin with, such as those which explicitly link the virus to China [89].

At the core of our findings, we underscore the embedded nature of online hate speech within large-scale interactional contexts and potential information maneuvers. In particular, we show that the relationship between bot activity and hate speech may not be straightforward on a tweet level, but must account for global as well as localized group structures. Comparing the US and Philippine cases, our measurement of network centrality metrics captured how hateful bots may orient toward distinct strategies or achieve varied levels of success in accruing influence and participating in the conversation more broadly. Context-specific distinctions notwithstanding, we emphasize two overarching insights gleaned from this analysis. First, bots—especially hateful bots—are influential and control major aspects of information flow in the pandemic conversation. Second, humans play a significant role in the spread of hate speech; recognition of organic participation in online toxicity will, thus, be vital for considering avenues for its mitigation [71].

Identity-based analysis of hate speech targets likewise showed the unique ways hate speech may be instrumentalized. Directing hate at particular racial or ethnic minorities—racial hate—may orient toward sowing conflict and deepening intergroup divisions. Meanwhile, directing hate at particular political leaders—political hate—may orient toward weakening trust in government institutions and extant leadership. Although united by a common overarching phenomenon, these two types of hate may also have distinct downstream real-world consequences. Recognizing the diverse nature of online hate, thus, helps to disentangle the distinct ways the pandemic exacerbates societal fractures. In the context of information operations, they also shed light on underlying strategic objectives which direct tactical maneuvers in cyber warfare [15, 22].

But while the above results highlight nuanced, separate mechanisms characterizing online hate speech, we also uncovered common, potentially more universal processes. By paying attention to community level dynamics, we systematically showed shared conditions for the increase of hateful talk in public discourse. In line with extant theories of extreme communication, we specifically demonstrated the joint associations of more isolated communication and inauthentic bot activity with greater toxicity [8, 64, 70]. These insights also compellingly emphasize that narrative-based and network-based information maneuvers be studied in a complementary fashion, as manipulation of information flow in this case appears to relate strongly with shifting toxicity and aggression in the online conversation.

From a methodological standpoint, we finally remark on the wide variety of tools used to further understand online hate speech dynamics. More specifically, we show how machine learning, network science methods, as well as insights from social scientific theory can be used to better understand online hate speech as it spreads ‘in the wild’ [45, 63, 76, 80]. Burgeoning literature on online disinformation and social cybersecurity advocates such integrated approaches to understanding sociotechnical problems like online hate speech [15, 22, 34]. This work illustrates how key principles of multidisciplinarity and interoperability may be applied in practice.

Toward global computational social science

The foregoing insights were facilitated specifically by adopting a global approach to computational social science [49]. By paying attention to multiple contexts, we situated our findings within a comparative frame, thereby allowing us to see which outcomes may be idiosyncratic to certain geopolitical settings, and which observed mechanisms may potentially be more general [79]. Furthermore, by situating online conversations within these geopolitical milieus, we were afforded more contextualized analysis.

For instance, observing higher levels of racial hate in the US, but not in the Philippines, made sense in relation to known societal conditions delineating the two nations [20, 37, 58]. While political polarization does indeed represent a serious problem for both nations [1, 8, 81], the multicultural setting of the US points to a fertile ground for racialized hate in a way that may not be comparable in the Philippines [5, 28, 44, 89]. These findings also usefully show the localized boundaries of universally framed claims about the rise of racist discourse in relation to COVID-19 [73]. Our work, thus, usefully joins burgeoning scholarship aiming to extend beyond predominantly WEIRD populations in the social sciences, computational and otherwise [39, 41]. As we hoped to demonstrate, this further strengthens our understanding of phenomena which are increasingly borderless yet may nonetheless feature unique dynamics within specific contexts.

Against the backdrop of the pandemic, both contexts face significant challenges at controlling outbreaks, and have extensively documented cases of online disinformation disrupting conversation in the public sphere [4, 9, 16, 62, 78, 88]. Our findings point to crucial evidence of social conflicts exacerbated under these conditions, both organically as well as inorganically. As researchers seek to understand the pandemic's impacts as an issue of both public health and societal importance more broadly, we suggest that these insights also matter for cultivating healthier digital ecosystems beyond the pandemic [15, 22, 55], as well as responding to genuine inequities which exist offline [10, 31, 51, 54]

Limitations and future work

Finally, this research features several limitations which likewise point to potential future work. First, we consider that our reliance on machine learning algorithms assumes comparability in the distributions of quantities we train our models on and the quantities which emerge in the real-world setting we examine here. Although we used relatively general training data and achieved performant experimental accuracy [13, 30], we note that COVID-19 represents an unprecedented phenomenon which may feature new and distinct patterns of public discourse. Thus, although we are confident in the broad patterns discerned in our analysis, individual predictions may feature errors unaccounted for given the novelty of the event being studied. Further work may aim to deepen the characterizations we offer here especially through qualitative assessment of predictions

in relation to empirical observations. As models continue to evolve to adapt to dynamic phenomena, future methods may also be improved through attention to these idiosyncrasies.

Second, we point out that in line with our global approach, more drill-down analysis may be warranted to deepen our understanding of online hate and disinformation as complex sociotechnical phenomena. Our findings on both network influence and identity targets not only capture systematic tendencies in the data, but they also point to more complex social processes which our methods are not designed to capture. Highlighting the multidisciplinary nature of social cybersecurity and computational social science more broadly, our work points to promising avenues for future work in this area. More specifically, psychological processes of responding to hate in the community-level context of inauthentic bot activity [17, 50] may be further probed in more controlled settings. Pivoting to the societal level of analysis, more macrosocial explanations regarding the differences in racially charged versus politically contentious hate in the US and the Philippines may also be explored [26, 54, 61].

Third, considering the large-scale nature of the global conversation about the pandemic, it may also be important for future work to consider the analysis shown here on more general datasets. Sampling in social media research can be fraught with generalizability issues [60]. Notwithstanding the limitations acknowledged in Sect. 3.1, our strategies for data collection may also feature key limitations by constraining our search parameters with respect to the broader COVID-19 conversation. The availability of open datasets for the COVID-19 discourse on Twitter, however, mitigates these issues by providing standard benchmarks for researchers studying the same phenomenon more universally [24]. Thus, research which assumes a less narrow theoretical focus than ours may fruitfully extend our analysis to these larger-scale datasets.

Acknowledgements This work was supported in part by the Knight Foundation and the Office of Naval Research Grants N000141812106 and N000141812108. Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS) and the Center for Informed Democracy and Social Cybersecurity (IDeaS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research or the U.S. government.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Abramowitz, A., & McCoy, J. (2019). United states: Racial resentment, negative partisanship, and polarization in trump's America. *The Annals of the American Academy of Political and Social Science*, *681*(1), 137–156.
2. Alorainy, W., Burnap, P., Liu, H., & Williams, M. L. (2019). The enemy among us: Detecting cyber hate speech with threats-based othering language embeddings. *ACM Transactions on the Web (TWEB)*, *13*(3), 1–26.
3. Antoci, A., Delfino, A., Paglieri, F., Panebianco, F., & Sabatini, F. (2016). Civility vs. incivility in online social interactions: An evolutionary approach. *PloS One*, *11*(11), e0164286.

4. Arif, A., Stewart, L. G., & Starbird, K. (2018). Acting the part: Examining information operations within #BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–27.
5. Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-muslim hate crime: Dynamics and impacts. *Aggression and Violent Behavior*, *27*, 1–8.
6. Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 258–265). IEEE.
7. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion* (pp. 759–760).
8. Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., et al. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221.
9. Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., et al. (2020). Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the National Academy of Sciences*, *117*(1), 243–250.
10. Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139.
11. Beskow, D. M. (2020). *Finding and characterizing information warfare campaigns*. Ph.D. thesis, Carnegie Mellon University.
12. Beskow, D. M., & Carley, K. M. (2018). Bot conversations are different: Leveraging network metrics for bot detection in Twitter. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 825–832). IEEE.
13. Beskow, D. M., & Carley, K. M. (2019). Agent based simulation of bot disinformation maneuvers in Twitter. In: *2019 Winter simulation conference (WSC)* (pp. 750–761). IEEE.
14. Beskow, D. M., & Carley, K. M. (2020). Characterization and comparison of Russian and Chinese disinformation campaigns. In *Disinformation, misinformation, and fake news in social media* (pp. 63–81). Springer.
15. Beskow, D., Carley, K. M.: *Social cybersecurity*. Springer **(forthcoming)**.
16. Beskow, D. M., & Carley, K. M. (2019). Social cybersecurity: An emerging national security requirement. *Military Review*, *99*(2), 117.
17. Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, *41*, 3–33.
18. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, É. (2011). The Louvain method for community detection in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *10*, P10008.
19. Borgatti, S. P., Carley, K. M., & Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, *28*(2), 124–136.
20. Bradshaw, S., & Howard, P. N. (2018). The global organization of social media disinformation campaigns. *Journal of International Affairs*, *71*(1.5), 23–32.
21. Calvert, C. (1997). Hate speech and its harms: A communication theory perspective. *Journal of Communication*, *47*(1), 4–19.
22. Carley, K. M., Cervone, G., Agarwal, N., & Liu, H. (2018). Social cyber-security. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation* (pp. 389–394). Springer.
23. Carley, L.R., Reminga, J., & Carley, K.M. (2018). Ora & netmapper. In: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer
24. Chen, E., Lerman, K., & Ferrara, E. (2020). COVID-19: The first public coronavirus Twitter dataset. arXiv preprint arXiv:2003.07372.
25. Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, *40*, 108–118.
26. Chiriboga, D., Garay, J., Buss, P., Madrigal, R. S., & Rispel, L. C. (2020). Health inequity during the COVID-19 pandemic: A cry for ethical global leadership. *The Lancet*, *395*(10238), 1690–1691.
27. Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811–824.

28. Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*, 1241.
29. Cresci, S., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2019). Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on twitter. *ACM Transactions on the Web (TWEB)*, *13*(2), 1–27.
30. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Eleventh international AAAI conference on web and social media*
31. Devakumar, D., Shannon, G., Bhopal, S. S., & Abubakar, I. (2020). Racism and discrimination in COVID-19 responses. *The Lancet*, *395*(10231), 1194.
32. ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth international AAAI conference on web and social media*
33. ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. In: *Twelfth international aaai conference on web and social media*
34. Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104.
35. Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 1–30.
36. Gallotti, R., Valle, F., Castaldo, N., Sacco, P., & De Domenico, M. (2020). Assessing the risks of" infodemics" in response to covid-19 epidemics. arXiv preprint arXiv:2004.03997.
37. Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2018). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In: *Proceedings of the 2018 World Wide Web Conference* (pp. 913–922).
38. Geiger, R. S. (2016). Bot-based collective blocklists in Twitter: The counterpublic moderation of harassment in a networked public space. *Information, Communication and Society*, *19*(6), 787–803.
39. Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not weird: The promise of the internet in reaching more diverse samples. *Behavioral and Brain Sciences*, *33*(2–3), 94.
40. Gunturi, V. M., Shekhar, S., Joseph, K., & Carley, K. M. (2017). Scalable computational techniques for centrality metrics on temporally detailed social network. *Machine Learning*, *106*(8), 1133–1169.
41. Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29.
42. Horton, R. (2020). Offline: COVID-19—What we can expect to come. *Lancet (London, England)*, *395*(10240), 1821.
43. Innes, M. (2020). Techniques of disinformation: Constructing and communicating "soft facts" after terrorism. *The British Journal of Sociology*, *71*(2), 284–299. https://doi.org/10.1111/1468-4446.12735._eprint.
44. Johnson, N. F., Leahy, R., Restrepo, N. J., Velasquez, N., Zheng, M., Manrique, P., et al. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, *573*(7773), 261–265. https://doi.org/10.1038/s41586-019-1494-7.
45. Joseph, K., Wei, W., Benigni, M., & Carley, K. M. (2016). A social-event based approach to sentiment analysis of identities and behaviors in text. *The Journal of Mathematical Sociology*, *40*(3), 137–166.
46. Kennedy, B., Jin, X., Davani, A. M., Dehghani, M., & Ren, X. (2020). Contextualizing hate speech classifiers with post-hoc explanation. arXiv preprint arXiv:2005.02439.
47. Kim, B. (2020). Effects of Social Grooming on Incivility in COVID-19. *Cyberpsychology, Behavior, and Social Networking*.,. https://doi.org/10.1089/cyber.2020.0201.
48. Krackhardt, D., & Stern, R.N. (1988). Informal networks and organizational crises: An experimental simulation. In *Social Psychology Quarterly*, 123–140
49. Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., et al. (2009). Life in the network: The coming age of computational social science. *Science (New York, NY)*, *323*(5915), 721.
50. Leader, T., Mullen, B., & Rice, D. (2009). Complexity and valence in ethnophaulisms and exclusion of ethnic out-groups: What puts the" hate" into hate speech? *Journal of Personality and Social Psychology*, *96*(1), 170.
51. Li, Y., & Galea, S. (2020). Racism and the COVID-19 epidemic: Recommendations for health care workers. *American Journal of Public Health*, *110*(7), 956–957.

52. Luengo-Oroz, M., Hoffmann Pham, K., Bullock, J., Kirkpatrick, R., Luccioni, A., Rubel, S., et al. (2020). Artificial intelligence cooperation to support the global response to COVID-19. *Nature Machine Intelligence*, *2*(6), 295–297. https://doi.org/10.1038/s42256-020-0184-3.
53. MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS One*, *14*(8), e0221152.
54. Martinez-Juarez, L.A., Sedas, A.C., Orcutt, M., & Bhopal, R. (2020). Governments and international institutions should urgently attend to the unjust disparities that COVID-19 is exposing and causing. *EClinicalMedicine*
55. Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., et al. (2019). Thou shalt not hate: Countering online hate speech. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 369–380.
56. Mohar, B. (1989). Isoperimetric numbers of graphs. *Journal of Combinatorial Theory, Series B*, *47*(3), 274–291.
57. Mønsted, B., Sapieżyński, P., Ferrara, E., & Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PloS One*, *12*(9), e0184148.
58. Montiel, C. J., Boller, A. J., Uyheng, J., & Espina, E. A. (2019). Narrative congruence between populist president Duterte and the Filipino public: Shifting global alliances from the United States to China. *Journal of Community and Applied Social Psychology*, *29*(6), 520–534.
59. Morgan, S. (2018). Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of Cyber Policy*, *3*(1), 39–43.
60. Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? comparing data from Twitter's streaming API with Twitter's firehose. In *Seventh international AAAI conference on web and social media*.
61. Ong, J. C., & Cabañes, J. V. A. (2018). Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines. Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines.
62. Ong, J. C., Tapsell, R., & Curato, N. (2019). *Tracking digital disinformation in the 2019 Philippine midterm election*. New Mandala.
63. Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547–577.
64. Pohjonen, M., & Udupa, S. (2017). Extreme speech online: An anthropological critique of hate speech debates. *International Journal of Communication*, *11*, 19.
65. Priante, A., Hiemstra, D., Van Den Broek, T., Saeed, A., Ehrenhard, M., & Need, A. (2016). # whoami in 160 characters? classifying social identities based on twitter profile descriptions. In: *Proceedings of the first workshop on NLP and computational social science* (pp. 55–65).
66. Reicher, S., & Stott, C. (2020). On order and disorder during the COVID-19 pandemic. *British Journal of Social Psychology*, *59*(3), 694–702.
67. Roussos, G., & Dovidio, J. F. (2018). Hate speech is in the eye of the beholder: The influence of racial attitudes and freedom of speech beliefs on perceptions of racially motivated threats of violence. *Social Psychological and Personality Science*, *9*(2), 176–185.
68. Rutledge, P. E. (2020). Trump, covid-19, and the war on expertise. *The American Review of Public Administration*, *50*(6–7), 505–511.
69. Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 1–9.
70. Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146.
71. Starbird, K. (2019). Disinformation's spread: Bots, trolls and all of us. *Nature*, *571*(7766), 449–450.
72. Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–26.
73. Stechemesser, A., Wenz, L., & Levermann, A. (2020). Corona crisis fuels racially profiled hate in social media networks. *EClinicalMedicine*,. https://doi.org/10.1016/j.eclinm.2020.100372.
74. Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, *115*(49), 12435–12440.
75. Stewart, L. G., Arif, A., Nied, A. C., Spiro, E. S., & Starbird, K. (2017). Drawing the lines of contention: Networked frame contests within #BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1–23.

76. Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.
77. Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 1–12.
78. Uyheng, J., & Carley, K. M. (2019). Characterizing bot networks on Twitter: An empirical analysis of contentious issues in the Asia-Pacific. In *International conference on social computing* (pp. 153–162). Behavioral-cultural modeling and prediction and behavior representation in modeling and simulation Washington DC, USA: Springer.
79. Uyheng, J., & Carley, K.M. (2020). Bot impacts on public sentiment and community structures: Comparative analysis of three elections in the Asia-Pacific. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*. Springer, Washington DC, USA.
80. Uyheng, J., & Montiel, C.J. Populist polarization in postcolonial Philippines: Sociolinguistic rifts in online drug war discourse. European Journal of Social Psychology (in press). https://doi.org/10.1002/ejsp.2716.
81. Uyheng, J., Magelinski, T., Villa-Cox, R., Sowa, C., & Carley, K. M. (2019). Interoperable pipelines for social cyber-security: Assessing Twitter information operations during NATO Trident Juncture 2018. *Computational and Mathematical Organization Theory*, 1–19.
82. Van Bavel, J.J., Baicker, K., Boggio, P.S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A.J., Douglas, K. M., & Druckman, J. N., et al. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, 1–12.
83. Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. In: *Eleventh international AAAI conference on web and social media*.
84. Waqas, A., Salminen, J., Jung, Sg, Almerekhi, H., & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PloS One*, *14*(9), e0222194.
85. 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).
86. Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2020). Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 93–117.
87. Woolley, S. C. (2016). Automating power: Social bot interference in global politics. First Monday.
88. World Health Organization: Coronavirus disease (COVID-19) weekly epidemiological update. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200914-weekly-epi-update-5.pdf (2020).
89. Ziems, C., He, B., Soni, S., & Kumar, S. (2020). Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. arXiv preprint arXiv:2005.12423.