Characterizing Online Engagement with Disinformation and Conspiracies in the 2020 U.S. Presidential Election

Karishma Sharma, Emilio Ferrara, Yan Liu
University of Southern California
krsharma@usc.edu, emiliofe@usc.edu, yanliu.cs@usc.edu,

Abstract

Identifying and characterizing disinformation in political discourse on social media is critical to ensure the integrity of elections and democratic processes around the world. Persistent manipulation of social media has resulted in increased concerns regarding the 2020 U.S. Presidential Election, due to its potential to influence individual opinions and social dynamics. In this work, we focus on the identification of distorted facts, in the form of unreliable and conspiratorial narratives in election-related tweets, to characterize discourse manipulation prior to the election. We apply a detection model to separate factual from unreliable (or conspiratorial) claims analyzing a dataset of 242 million election-related tweets. The identified claims are used to investigate targeted topics of disinformation, and conspiracy groups, most notably the far-right QAnon conspiracy group. Further, we characterize account engagements with unreliable and conspiracy tweets, and with the QAnon conspiracy group, by political leaning and tweet types. Finally, using a regression discontinuity design, we investigate whether Twitter’s actions to curb QAnon activity on the platform were effective, and how QAnon accounts adapt to Twitter’s restrictions.

Introduction

Disinformation and social media manipulation has threatened the integrity of elections and democracies around the world. In the 2016 U.S. Presidential Election, official investigations by the U.S. Congress revealed the presence of Russian state-backed operations attempting to manipulate the U.S. political landscape. In the years following the investigation, social media platforms like Twitter and Facebook have actively documented several other suspicious operations linked to Russia, Iran, Venezuela, China and other countries (Gadde and Roth 2018), suggesting presence of persistent efforts to manipulate online discussions. Although social media has been instrumental in shaping conversations around social and political issues (Loader and Mercea 2011), its positive effects have been undermined by disinformation and manipulation. The promotion of disinformation, propaganda, and politically divisive narratives have been regularly seen in many social contexts (Sharma, Ferrara, and Liu 2020; Bessi and Ferrara 2016; Martin and Shapiro 2019).

Distortion of facts to promote disinformation and propaganda greatly reduces trust in online systems, because of its ability to influence both individual opinions and social dynamics (Van der Linden 2015; Nyhan and Reifler 2010). At an individual level, exposure to unreliable and conspiratorial narratives has been observed to influence people’s perceptions of the truth and decrease pro-social behaviours (Van der Linden 2015; Nyhan and Reifler 2010). Moreover, with polarized and partisan narratives, it can further exacerbate prejudices and ideological separations (Jolley, Meleady, and Douglas 2020). Secondly, the disinformation landscape is complex, with possibilities for injecting disinformation, propaganda, and influencing the discourse through automation (Ferrara et al. 2020), coordination (Sharma, Ferrara, and Liu 2020), compromised accounts, follow trains and others (Torres-Lugo, Yang, and Menczer 2020). The range of activities can be widespread, with both concealed and overt efforts to promote manipulative narratives.

In the context of the 2020 U.S. 2020 presidential election, held on November 3, 2020, since Twitter is recognized as one of the social media platforms with the most news-focused users, with significant following in U.S. politics (Hughes and Wojcik 2019 (accessed March 20, 2020), we investigate and characterize disinformation on Twitter, with an analysis of more than 240 million election-related tweets, collected between June 20, 2020 and September 6, 2020. We focus on the following research questions to characterize online engagement with disinformation and conspiracies:

- **R1.** What are the prevalent disinformation and conspiracy narratives on Twitter preceding the U.S. 2020 Election? We apply and train a detection model to separate factual from unreliable (or conspiratorial) claims to examine how the political discourse was manipulated.
- **R2.** How significant is the impact and reach of disinformation and conspiracy groups in terms of account engagements (characterized by activity level, political leaning, tweet types, and propagation dynamics)? We characterize account engagements with the QAnon conspiracy group, and we compare propagation dynamics of unreliable/conspiracy and reliable engagement cascades.
- **R3.** Did Twitter’s restrictions on QAnon influence its activities and were they effective in limiting the conspiracy? We investigate activity before/after the restric-
tion and factors driving the sustained activity of QAnon accounts, and we test if QAnon accounts adapted to Twitter’s restrictions using a regression discontinuity design.

**Contributions and findings.** The findings of the research questions studied in this paper are as follows:

- We apply CSI (Ruchansky, Seo, and Liu 2017) to detect unreliable/conspiracy tweets in the election dataset, with validation AUC 0.8 and F1 0.76. The most prominent topics targeted by disinformation promoters preceding the election included mail-in voter fraud, COVID-19, Black Lives Matter, media censorship, and false claims and conspiracies about past and current political candidates.

- While disinformation tweets are widespread and diverse, engagements with such tweets are less viral than with reliable tweets (i.e., mean time to reach unique accounts is higher, mean cascade breath is smaller). Yet, there can be several such tweets which manage to receive substantial attention and engagements.

- QAnon far-right conspiracies are quite prevalent on Twitter, and we find that while 85.43% of all accounts (10.4M) in the collected data, did not have an observed engagement with QAnon accounts in the dataset, fraction of such accounts is much smaller (34.42%) in the active 1.2 million accounts (that appear in at least 20 tweets in data).

- Investigation of QAnon activities suggests that their strategies include entering into active discussions with left-leaning accounts through replies, in an effort to “red pill” (i.e., dramatically transform ones perspective with typically disturbing revelations). Also, the conspiracy group quickly adapted to Twitter’s ban on its activities imposed in July, 2020. Using a regression discontinuity design to estimate causal effects of Twitter’s intervention, we find statistically significant changes in hashtags adopted by QAnon. Also, the volume of engagements after the Twitter ban is still largely sustained by existing accounts created before the ban, rather than new accounts.

**Research implications.** Our findings indicate that there are high interactions among more active accounts and disinformation or conspiracy groups, and engagements through reply tweets are prevalent with both left/right leaning accounts. An important implication from the study is also that the actions enforced to limit malicious groups might not be robust to evasive actions from malicious actors. Therefore, rethinking effective platform interventions is needed.

This work is different from most previous studies, especially in investigating effectiveness of social media platform actions to limit disinformation or conspiracies. QAnon is a notable conspiracy, and most studies have traced its normalization, its conspiracies and posts. Our work instead is the first to focus on characterization of interactions with QAnon accounts, and their activities and evasion of Twitter’s restrictions. We also provide a large data-driven analysis of disinformation preceding the 2020 U.S. Election, focused on observed engagements for quantifying the impact of disinformation/conspiracies to better understand its risks.

**Related Work**

Social media manipulation has been observed in political discourse in several countries (Woolley and Howard 2017). However, changing strategies of social media manipulation continue to challenge our understanding and findings on malicious activities on the network (Cresci 2020). Twitter reported seeing some of the operations as being less clandestine and more overt efforts to promote disinformation than they have in recent years (Miles P. 2020 [accessed June 18, 2020]). Therefore, identification and analysis of disinformation and its landscape preceding the U.S. 2020 election is vital, along with characterization of account engagements with disinformation, to estimate its risks, and investigate the effectiveness of interventions and actions to limit it.

With regards to U.S. elections, recent papers attempt to uncover different aspects of social media manipulation. (Bessi and Ferrara 2016) discovered tens of thousands of bot accounts distorting the 2016 election-related Twitter discourse, in what turned out to be an effort associated with the Russian Internet Research Agency (Badawy et al. 2019). (Luceri, Cardoso, and Giordano 2020) conducted a year-long analysis of bot activity on Twitter before the 2018 midterm election. Based on detection of bot-like behaviours in accounts, they found that 33% of bots were active even one year prior to the midterm, and new bots were introduced into the discussion progressively up to the midterm. Nineteen thousand highly-active bots generated a large volume of 30M political messages, and humans were susceptible to retweeting bots with one-third of their retweets being content pushed by bots. (Ferrara et al. 2020) observed high volume of tweets from bot-like accounts prior to the November election, and found highly partisan retweeting behavior among bots and humans. (Ferrara et al. 2020) also analyzed distortion with three types of conspiracy hashtags, i.e. QAnon, -gate conspiracies such as pizzagate and obamagate, and COVID-19 conspiracies, and found 13% of all accounts sharing the conspiracy hashtags were suspected bots. They also found that accounts sharing content from right-leaning media were almost 12 times more likely to share conspiratorial narratives than left-leaning media accounts.

Tracking the growth of the QAnon conspiracy from fringe online subcultures to mainstream media over one year. (de Zeeuw et al. 2020) found that after the incubation period on 4chan/pol in Oct 2017, the movement had quickly migrated to larger platforms, notably YouTube and Reddit, and was covered by news media only when it moved off-line. In terms of cross-platforms effects, (Papasavva et al. 2020) found spike in new account registrations on Voat after Reddit ban on QAnon threads in Sept 2018. They also studied content of QAnon posts finding discussions related to predecessor Pizzagate conspiracy, and “Q” drops, which refer to the cryptic posts meant for adherents of the conspiracy to decode, and found more submissions but less comments in QAnon threads compared to general threads.

Addressing growing concerns around conspiracy communities. (Phadke, Samory, and Mitra 2021) investigated what individual and social factors promote users to join conspiracy communities. Their findings on Reddit suggest that direct interactions (i.e., replies) with conspiracists is the most
important social precursor for joining (i.e., first contribution to) a conspiracy community. (Silva et al. 2020), on the other hand, investigated which factors are most predictive of engagements with factual and misleading tweets, finding follower-friend ratio, banner image or URL in user profile, presence of image in the tweet, as most relevant, and factual tweets were more engaging in COVID-19 tweets. Differently from the papers, we focus on characterization of interactions with QAnon accounts, based on political leaning and tweet types, and on their activities and strategies to engage with other accounts and escape Twitter’s restrictions.

Data Collection

For the analysis, we start from the dataset collected in (Chen, Deb, and Ferrara 2020) which tracks election related tweets from May, 2019 onwards, and contains over approximately one billion tweets. The dataset was collected by tracking mentions of official and personal accounts of Republican and Democratic candidates in the presidential election using Twitter’s streaming API service, which returns matches in tweet content and metadata within a ~1% sample of the stream of all tweets. The details of the tracked mentions and distribution of frequent hashtags and bigrams in the data are available in (Chen, Deb, and Ferrara 2020).

For this analysis, we focus on the tweets that appeared between June 20, 2020 and September 6, 2020, in order to study the disinformation surfaced on Twitter in the months preceding the election. This subset of the data contains 242,087,331 election-related tweets from 10,392,492 unique users. We focus on engagements, therefore, it is useful to define four tweet types considered here: (i) Original tweets (accounts can create content and post on Twitter) (ii) reply tweets (iii) retweeted tweets, which reshare without comment (iv) quote tweets (embed a tweet i.e., reshare with comment). Engagements are defined here as one of the tweet types other than original tweets (and engagements can be with original tweets or other tweet types also).

Disinformation Detection Methodology

In this section, we present the methodology to separate factual from unreliable (or conspiratorial) claims in election-related tweets. Conspiracies theories are attempts to explain events or situations, often postulated on most likely false and unverifiable theories, and may be politically motivated (Ferrara et al. 2020; Phadke, Samory, and Mitra 2021). To capture disinformation aimed at manipulation of public opinion, we focus on any kind of unreliable (likely false or misleading) narratives, including conspiracies promoted prior to the election. The term disinformation is used here as an umbrella term to refer to distorted narratives appearing in the form of unreliable, including conspiratorial claims. In literature, the term disinformation separates itself from misinformation based on intent to manipulate vs. lack thereof (Sharma et al. 2019). However, since we do not assess intent from tweets, we do not make a distinction between the two terms.

Model Specifications

Detection model. We apply CSI, a supervised deep learning based model in (Ruchansky, Seo, and Liu 2017) for disinformation detection. The model captures three main signals or features: source, text, temporal, which are useful for disinformation detection (Ruchansky, Seo, and Liu 2017; Sharma et al. 2019). To classify a tweet, the model considers time-ordered sequence of engagements that the tweet receives, along with temporal information such as the frequency and time intervals between engagements. In addition, it learns the source characteristic, based on account behaviours i.e., which accounts co-engage with a given tweet. These input signals are useful for differentiating disinformation tweets. In general, engagements or social context are useful to learn predictive features, as they provide feedback from other accounts through engagement dynamics (Sharma et al. 2019). The model needs to be provided training data containing labeled cascades (i.e., tweets with its engagements). We describe the process for labeling data and extracting features (cascades) for training the detection model in the following paragraphs.

Reliable, unreliable and conspiracy URLs. We collect lists of reliable and conspiracy news sources from three fact-checking resources on low-credibility news sources: Media Bias/Fact Check, NewsGuard, and Zimbars (2016). NewsGuard maintains a repository of news publishing sources that have actively published false information about the recent COVID-19 pandemic. The listed sources from NewsGuard, accessed on September 22, 2020 are included, along with low and very low factual sources listed as questionable from Media Bias/Fact Check, and sources tagged with unreliable or related labels from Zimdar’s list. We separately collect mainstream reliable news sources referencing Wikipedia. In total, we obtained 124 mainstream reliable and 1380 unreliable (or conspiracy) news sources. Original tweets sharing URLs published from these news sources are thereby labeled as reliable and unreliable/conspiracy respectively. In addition, we also label tweets that are retweets, replies, or quotes sharing such URLs, if the parent of the tweet is not in the collected tweets (1% Twitter sample).

Table 1: Detection dataset statistics

| Statistic                          | Count       |
|-----------------------------------|-------------|
| Tweets                            | 242,087,331 |
| Users accounts                    | 10,392,492  |
| Unreliable/conspiracy URLs cascades | 3,162      |
| Reliable URLs cascades            | 4,320       |
| Unlabeled cascades                | 192,103     |
| Avg. cascade size (# engagements) | 57.11       |
| Avg. cascade time (in hrs)        | 80.42       |
| Avg. time between engagements (hrs)| 5.93        |
Information cascades and cascade statistics. To train the detection model, we first extract cascades of engagements from the dataset. Information cascades represent the diffusion or spread of information on the network. Each cascade corresponds to a time-ordered sequence of engagements (retweets, quotes, and replies), that originate at a source tweet and spread through chains of retweets, quotes and replies. Formally, a cascade can be denoted as $C_j = [(u_1, tw_1, t_1), (u_2, tw_2, t_2), \cdots (u_n, tw_n, t_n)]$, where the tuple corresponds to an original tweet, or an engagement with the original tweet or its predecessor engagements, and represents the user ($u$), tweet ($tw$), and temporal ($t$) dimensions of when the user posted the tweet or engagement. The cascade is labeled by the URLs linked in its first tweet as described earlier, if the tweet shares a URL from one of the considered factual or unreliable news sources, it is labeled accordingly, otherwise left unlabeled.

Since the dataset contains over 10M accounts and 242M tweets, for computational efficiency, we subsample accounts and cascades, whilst ensuring minimal information loss, such that at least 75% (~180M) of the tweets are accounted for in the data. We find that this can be achieved by subsampling cascades with a minimum number of engagements, and accounts with highest active and passive engagements with other accounts in the dataset. For accounting for >75% of the tweets, we find cascade size of 5, and 7471 highest engaging accounts was sufficient. This results in ~200K cascades (192K unlabeled, and rest labeled with 3120 unreliable and 4320 reliable cascades). The full cascades comprise engagements from accounts outside of the subsampled accounts, and span the large fraction of dataset tweets. Full cascades are used in the analysis of identified disinformation cascades, but for training and inference with CSI, the subsampled accounts in these cascades are utilized. Table 1 provides statistics of data and training cascades.

Model Evaluation

Table 2 reports 5-fold cross validation results obtained using CSI on the labeled cascades. The metrics reported are the ROC AUC average precision AP score. We also reported the F1, Precision, Recall and macro-F1 at the detection threshold selected from the ROC curve, that achieves maximum geometric mean of validation set sensitivity and specificity.

For comparison, we evaluate it against several baselines: (i) SVM- RBF on text features of the cascade source tweet, extracted with doc2vec. (ii) GRU [Ma et al. 2016] which utilizes a recurrent neural network (RNN) to classify the cascade based on text features of the time-ordered sequence of engagements in the cascade. (iii) CI, which can be considered as a variant of the CSI model, utilizes only tweet text and account metadata features of the engagements as input to a RNN for classification. (iv) CI-t, another variant of the CSI model, utilizes tweet features of engagements, and in addition temporal information of time intervals between engagements as input to a RNN. (v) CSI model, which includes all three features (tweet features of engagements, temporal information, and account behaviours).

The CSI model has AUC 0.8 and high recall and is therefore used for inference on unlabeled cascades. The ensemble of CSI models trained over five folds are used for inference. We take the top 80th and bottom 80% percentile, of cascades predicted with the largest margins above and below the detection threshold, resulting in 72,228 unreliable and 81,453 reliable classified cascades respectively. The engagements or tweets forming the reliable cascades constitute about two third of the total tweets in all cascades. For human validation, we take a random sample of 50 cascades from ones labeled by the model (Note: the list of validated tweets is provided in supplementary materials). In 2/50 (4%) the model label differed from the label assigned by inspecting the tweet and its account for suspensions. The model errors included a tweet critical of the postal system and though non conspiratorial, could have been mistaken for claims related to mail-in voter fraud by the model, and another which strongly supported criticism of China by president Trump, although not from a suspended account and unlikely part of QAnon/conspiracy groups, was predicted unreliable. To reiterate, these validated cascades were uniformly sampled from cascades that were assigned labels using CSI, considering predictions with largest margins from the detection threshold, as discussed earlier using the ensemble of 5-fold classifiers, and the expected lower error rate, confirmed by validation suggests it can be reliably utilized for further analysis.

Methodology for Inferring Political Leaning

In this section, we describe the methodology we use to infer political leaning of accounts on the network. The inferred political leanings are used later in the analysis and characterization of the disinformation landscape. Similar to prior work [Ferrara et al. 2020], we use the list of 29 prominent news outlets classified as left, lean left, center, lean right, right as per ratings provided by allsides.com We consider...
left and lean left classified outlets as the left-leaning, and right and lean right as right-leaning.

Model specifications
For accounts that tweeted or retweeted URLs published from these news outlets, we characterize their political leaning by measuring the average bias of the media outlets they endorsed. This gives us a set of labeled accounts with known political leaning labels based on the media URLs. To infer labels of other accounts, we can propagate labels from this known set (also called the seed set i.e., accounts with known political leaning labels based on the media URLs) to other accounts based on interactions between accounts.

We utilize the retweet network to infer political leaning of other accounts starting from the seed set accounts. Retweeting is a form of endorsement, different from other tweet types, and accounts which retweet each other tend to share the same political biases (Badawy et al. 2019). To that end, we use Louvain method (Blondel et al. 2008) to identify communities in the retweet graph, where edge weights represent the number of retweets between accounts. The method optimizes modularity which provides a measure of edge density within communities compared to across communities. We assign political leaning to each identified community, using the average political leaning of media URLs endorsed by accounts in the seed set that belong to the community. The seed set accounts with high entropy in distribution of left-leaning and right-leaning URLs (close to uniform distribution with margin of 0.2), and ones that shared less than 10 URLs from the media outlets are filtered out.

Inference of political leanings
Using media outlets, we obtain a seed set of 114K accounts from 10.4M accounts in the dataset. To limit the size of the retweet network, we consider the top active 1.2M accounts that appeared (in original tweet, retweet, quote, or reply tweet) at least 20 times in the collected dataset.

Using the seed set and retweet graph of 1.2M accounts, gives a resulting inferred network with large left-leaning communities of 540,719 and 68,197 accounts and two smaller left-leaning ones, and large right-leaning communities of 480,982 and 10,723 accounts, and multiple smaller ones. Of the 1.2M accounts, we were able to infer the political leaning of 92% of the accounts. The rest of the accounts remain undetermined due to high entropy in left and right-leaning URLs shared, or with communities that had fewer than two seed accounts. We thereby identified 610,430 left-leaning URLs shared, or with communities that had fewer than two seed accounts. We thereby identified 610,430 left-leaning ones, and large right-leaning communities of 540,719 and 68,197 accounts and two

Table 3: Number of accounts labeled as left or right-leaning (by media URLs, account profile description, and human verification) for validation, with error rate (%) in each type based on the inferred political leaning of those accounts.

|                   | MediaURLs | Profile Desc. | Human Verif. |
|-------------------|-----------|---------------|--------------|
| LP-Left           | 68k (0.71)| 29.5k (0.32)  | 116 (1.72)   |
| LP-Right          | 46k (0.27)| 14.0k (0.34)  | 103 (2.91)   |
| Lo-Left           | 68k (0.67)| 29.5k (0.25)  | 116 (1.72)   |
| Lo-Right          | 46k (0.37)| 14.0k (0.58)  | 103 (2.91)   |

hashtag were classified as left/right through human validation of most frequently used 3,000 hashtags. The list is provided in supplementary material). (iii) Manual verification based on inspection of tweets of randomly sampled subset of accounts (i.e., based on explicitly stated party affiliation in tweet or account profile, or expressed support of left/right presidential candidate/party, and assigned ‘center’ instead of left/right if non-partisan). We sampled 100 inferred accounts uniformly at random, and 124 inferred accounts by stratified sampling based on degree distribution of accounts in the retweet graph, to ensure coverage of dense and sparsely connected accounts in the retweet graph, since the retweet graph was used infer the leanings.

In Table 3 we report the error rate on each evaluation measure separately for the left-leaning and right-leaning accounts. Here, we added an alternative baseline based on Label propagation (Badawy et al. 2019) for comparison. The total error rate was 4.46% on manually verified labels for both methods (label propagation (LP) and Louvain (Lo) used here, on RT graph from same media URL labeled seed set) and results were robust on both. Error analysis on the manually labeled set, suggests that errors included few accounts that are actually neither left/right leaning (but center), e.g., U.S. Department of State, reporter accounts, or unrelated or disinterested in U.S. politics, and were erroneously classified as left/right leaning. Other mistakes included anti-Trump conservatives that were inferred as left-leaning, or with high entropy in left/right leaning views in their tweets.

Results and Analysis
In the following subsections, we analyze identified unreliable/conspiracy cascades, and characterize account engagements. We characterize disinformation topics, QAnon conspiracy group and its interactions with left and right demographics, their activities, effect of Twitter’s restrictions, and propagation dynamics of engagements with disinformation.

Disinformation topic modeling
Topic modeling We use topic modeling to identify prominent topics that were targets of disinformation prior to the election. With identified unreliable/conspiracy cascades, we can model the topics in tweet text associated with the source (first) tweet in the cascade. The text is pre-processed by tokenization, punctuation removal, stop-word removal, and removal of URLs, hashtags, mentions, and special characters, and represented using pre-trained fastText word embeddings.
Get this straight America. Before COVID struck @realDonaldTrump facilitated record economic growth. Best unemployment rates for every demographic in 58 yrs. Manufacturing now what Obama said it would be impossible. He’s answered every call to fight COVID. Plus more. RT

Mail-in voting is a proven danger WE CAN LOSE AMERICA! The left is using the virus to push mail voting because they can’t win a fair election. Trump is more popular than ever. People are waking up. DO NOT GIVE UP FIGHT FOR OUR FUTURE: #2020Election

On November 3, 2020, the US has the historic opportunity to bury the Democrats Party. They promoted anarchy, China’s agenda, and hatred among Americans. Only four more years of Trump can guarantee security and progress for the US. @LindseyGApplies @CR2AZ @realDonaldTrump #VF

Jobs, tax, economy, us, china, American, people, federal, law, mining, taxes, order, need, would, state

On this Thursday, the 16th of July 2020, I want to say Thank You to @realDonaldTrump for his strength and love for our Country. God knew we needed your fight and perseverance to protect us against enemies abroad and near. God Bless you, President Trump!

god, bless, loveoneanother, bebest, good, patriots, morning, lord, pray, please, thank, participate, retweet, thanks,

Most of us who felt strongly about HCQ as a therapeutic suspected this @realDonaldTrump @realDonaldTrump. DEADLY COVER UP: Fauci Approved Hydroxychloroquine 15 Years Ago to Cure Coronavirus. "Nobody Needed to Die"

covid, virus, vaccine, coronavirus, aids, people, hydroxychloroquine, deaths, fda, treatment, hcq, china, patients, fauci

news, watch, twitter, media, video, fake, conference, fox, cnn, people, live, breaking, press, see, tweets

O Watch Missile Attack on 4th Sec @realDonaldTrump Was Right

q, maga, trust, tammy, biden, plan, bbb, rd, love, must, great, bqqm

police, bmm, law, antifa, violence, federal, riots, people, cities, democrats, terrorists, violent, crime, terrorist, stop

VENEZUELA CONFRONTATION
LCM "Venezuela and opposition's display of strength is nothing to be afraid of. We're here to stay." ENGLISH

venezuela, por, de, para, urgente, militar, lea, en, necesita, libertad, intervencion, socialistas

facebook.com/GoogleYouTube/Twitter Conor Viral Video of Decision

Coup in Hill Coronavirus Press Conference They all have Blood orWhite Will BLM Be Declared a Terror Organization and Be Imprisoned? It's Right for their hands because they're KNOWINGLY suppressing informationThe NZ TV Central Streets of These Dem-Ran Cities @realDonaldTrump Families that could save lives VOTE@realDonaldTrump

Reportedly Trapped in Bucking Set on Fire By Black Lives Matter Rioters in Kenosha

Figure 1: Topic clusters for identified unreliable/ conspiracy tweets with example top representative tweet of cluster

(Bojanowski et al. 2017)† We take the average of word embeddings in the tweet text to represent each tweet. Using pre-trained embeddings trained on large English corpora, we can potentially encode more semantic information and it is useful for short texts where word co-occurrence statistics are limited for utilizing traditional probabilistic topic models (Liu et al. 2016). The tweet text representations are clustered using k-means to identify topic clusters.

We select number of clusters (K) using silhouette and Davies-Bouldin measures of cluster separability. K that jointly is in the best scores of both is selected between 3-35. This gave us K=30, and inspecting word distribution and representative tweets (closest to cluster center), we discard two clusters unrelated to US politics (ALS treatment and Nigeria violence), and eight small or less distinguished clusters, and merge over-partitioned clusters each related to the Black Lives protests, and to mail-in voter fraud.

Topic clusters The resultant clusters are in Fig 1 For each cluster, top words ordered by highest tf-idf scores, along with an example tweet from 100 most representative tweets of the cluster are shown. The major themes relate to false claims about mail-in voter fraud, COVID-19 and pushing hydroxychloroquine as a cure, and protests concerning law enforcement and Black Lives Matter. Other topics target specific candidates and entities, such as social media platforms for censorship of unverified and conspiratorial content, controversies and allegations against former president Obama, or targeting the democratic party as a whole on different social issues, and misleading claims about jobs and economy. The remaining clusters include the QAnon conspiracies, a far-right conspiracy group now banned by several platforms (de Zeeuw et al. 2020). Another cluster related to Venezuela appears in support of right-leaning theories potentially about voter fraud, and anti-democratic posts, however, our modeling is limited to English, the most prominent language in the tweets (~94% disinformation tweets were in English, followed by ~3% in Spanish, remaining languages less than 0.1% each). In Table 4 we list examples of identified unreliable/conspiracy tweets within tweets with the most engagements in the collected dataset, discarding false positives. Some of these have been debunked by fact-checking sites as false, misleading or lacking evidence.

Quantifying interactions with QAnon group

QAnon, a far-right conspiracy group emerged in 2017 on 4chan, and has risen to prominence for its baseless conspiracies that have received significant following and attention (de Zeeuw et al. 2020). The group advances the conspiracy theory that president Trump is battling a satanic child sex-trafficking ring, and an anonymous ‘Q’ claims to be a US government official with top-clearance, providing insider information about deep state operations (Ferrara et al. 2020). The cluster advances the conspiracy theory that president Trump is battling a satanic child sex-trafficking ring, and an anonymous ‘Q’ claims to be a US government official with top-clearance, providing insider information about deep state operations (Ferrara et al. 2020). In this subsection, we analyze activities of QAnon accounts and characterize its interactions with other accounts.

Identification and verification. We identify accounts that actively form part of the conspiracy group by posting original content related to QAnon conspiracies, referred to as QAnon accounts thereafter. We extract tweets (original, https://www.politifact.com/factchecks/2020/jul/28/viral-image/opening-case-file-does-not-mean-joe-biden-criminal/)

https://apnews.com/article/shootings-wisconsin-race-and-ethnicity-politics-lifestyle-23668668b3b59fa609a18d023c0bb485

†Pre-trained: https://fasttext.cc/docs/en/english-vectors.html
Table 4: Examples of unreliable/conspiracy tweets within tweets with most engagements in the data

| Unreliable/conspiracy tweet                                                                 | # Eng. |
|---------------------------------------------------------------------------------------------|--------|
| BREAKING: Democratic Presidential nominee @JoeBiden is formally being listed as a criminal suspect by high level Ukraine government officials, in a major case involving his son - Hunter. https://t.co/Xe2sLEAh8 | 56K    |
| We are being censored! @realDonaldTrump @Facebook is about to unpublish our FB page with 6 million followers. The NY Times recent article claiming we are right Wing Provacateurs They are interfering with this election! Conservatives are being censored on FB. PLEASE RETWEET! https://t.co/xy8yZ7kyc | 48.6K  |
| I am tired of the censorship! Anderson Cooper called me a snake oil salesman because I'm trying to get the FDA to test a supplement that I've seen work! And Twitter keeps taking my followers! Please RT and follow me as I support President @realDonaldTrump? | 48K    |
| Can you believe what’s happening!? They give Joe Hiden’ the questions, and he reads them an answer! https://t.co/sfVMw6uQ2g | 47K    |

HEARTBREAKING A 60 year old Black Trump Supporter was murdered in cold blood all because he support President @realDonaldTrump This is a Hate Crime He deserves Justice Let's make his name trend Use #JusticeForBernellTrammell https://t.co/XZbdOiHgRR | 46.9K  |

NATURE ARTICLE HOAX BUSTED!! Proof that chloroquine let's covid attack cancer cells but not normal cells. PLEASE RETWEET. @realDonaldTrump @IngrahamAngle @SteveFDA @drsmonegoldog @jennybethin https://t.co/XN0YCIhiSQ | 40.6K  |

If we can stand in line at a grocery store or hardware store, we can stand in line at the polls to vote. President @realDonaldTrump is RIGHT that universal, unmonitored vote-by-mail would be a DISASTER, and we're already seeing evidence of that across the country. @TeamTrump https://t.co/aIUkJ7kQ | 32.5K  |

| Keyword                      | Freq.   | Keyword                       | Freq.   |
|------------------------------|---------|-------------------------------|---------|
| wwg1wga                      | 159436  | wgwolrdwide                   | 18231   |
| #qanon                       | 68039   | #garmy                        | 13577   |
| #obamagate                   | 78574   | #pizzagate                    | 13053   |
| #savethechildren             | 33221   | #takerhearth                  | 10994   |
| thegreatawakening             | 23305   | greateawakening                | 31615   |
| deepstate                    | 25268   | deepstatecup                  | 995     |
| deepstatecibal               | 1669    | deepstateexposed              | 2188    |
| #pedogate                    | 5211    | pedowood                      | 4454    |
| #plandemic                   | 8456    | #scandemc                     | 4674    |
| #sheepnomore                 | 1492    | adrenochrome                  | 6397    |
| thestom                    | 3989    | followthewhiterabbit          | 95      |
| thesepeolearesick           | 2843    | wearetheneuwsnow               | 5540    |
| trusttheplan                | 2579    | pizzagateresal                | 698     |
| themorsomousponus            | 897     | newworldorder                 | 901     |
| darktolight                   | 6825    | clintonbodycount              | 1898    |

Table 5: QAnon conspiracy keywords along with their occurrence frequency in tweets (original, reply or quoted tweets i.e., excluding retweets) containing the keywords.

Table 6: Verification of 100 accounts sampled from inferred right/left-leaning accounts posting QAnon associated keywords. Verified as: (Q) is QAnon conspirator; else not (N) words are not considered in the recall measure).

**QAnon interaction graph.** To quantify interactions between accounts, we consider that two accounts have interacted if there exists a retweet (i.e., retweet without comment), quoted tweet (i.e., retweet with comment), or reply tweet edge between the accounts, in the collected dataset. Further, to account for directionality of the interactions, we partition the accounts that have interacted with QAnon accounts into “ID” (influenced) and “IR” (influencer). Influenced accounts are those that have been influenced by (retweeted, quoted, or replied to) QAnon accounts tweets, whereas influencer are accounts that QAnon accounts have retweeted, quoted or replied to. Table 7 reports what percentage of all accounts were influenced but did not influence QAnon (ID-only), influencer but not influenced (IR-only), and both influencer and influenced in the data (IR&ID).

We find that 85.43% of 10.4M accounts had no observed interaction with QAnon accounts in the collected tweets. However, considering accounts that appeared in at least 20 tweets (original, reply, retweet, quote) in the dataset, we find only 34.42% of the active 1.2M accounts had no interaction with QAnon accounts in the collected tweets. Also, we consider a random sample of accounts from the active 1.2M accounts of the same sample size as % of QAnon accounts in that set (drawn five times, and statistics averaged) for comparison. IR&ID and IR-only are similar. The main difference is in the ID-only and Neither cases, wherein % of influenced...
The interaction graph of the 1.2M active accounts (activity > 20 in the dataset) is visualized in Fig. 2. Thus, 5.2% of these accounts are QAnon accounts, as described earlier. And 24.9% (IR&ID) are closely coupled accounts that both influenced QAnon content and were influenced by it (retweeted, quoted, replied). A larger 30.98% were only influenced (ID-only), while a small fraction of 4.5% were retweeted, quoted, or replied by QAnon, but did not engage back (IR-only). The edge weights in the interaction graph correspond to the volume of retweets, replies, or quoted tweets of the destination node by the source node, per account in the source node.

**QAnon engagement characteristics.** Next, we characterize interactions with QAnon accounts by inferred political leaning and tweet types, in the active 1.2M accounts in Fig. 3a. As one would expect, the retweeted and quoted tweet interactions are mostly between right-leaning and QAnon accounts (90.06% and 73.44%), as these are more close to endorsements. Replies on the other hand are roughly equally distributed amongst left and right-leaning accounts (49.97% of replies by QAnon are towards left-leaning and 42.54% of the replies to QAnon tweets are from left-leaning accounts).

This suggests that QAnon get engaged in active discussions with left-leaning accounts. This might also be indicative of engagement strategies by QAnon to influence liberals by “red pilling”, i.e., have their perspective transformed, illustrated in a QAnon account tweet which states:

> “We interact with liberals the most under realDonaldTrump’s most recent posts. Q gave us a challenge so I have been sharing the truths about COVID & nursing home deaths in hopes of redpilling anyone who sees. Next Trump tweet wanna help out? [Link to Twitter thread]”

Fig. 3b shows distribution of inferred left/right-leaning accounts that interacted with QAnon, by tweet types. 82.17% of left-leaning in 1.2M have not retweeted/quoted QAnon accounts, but only 28.35% of right-leaning have not endorsed content from QAnon accounts, but only 28.35% of right-leaning have not endorsed content from QAnon accounts.

### Estimating change in QAnon activities after Twitter Action

Twitter announced sweeping actions against the QAnon conspiracy group on July 21, 2020, designating them as harmful coordinated activity. Twitter took down more than 7,000 QAnon accounts violating platform manipulation policies with spam, ban evasion, and operation of multiple accounts. Twitter also banned QAnon-related terms and urls from ap-
pearing in trending topics and search, limiting 150,000.\footnote{https://www.nbcnews.com/tech/tech-news/twitter-bans-7-000-qanon-accounts-limits-150-000-others-n1234541}

**QAnon activities timeline.** Timeline of tweet volume (original, retweet, quote, reply) from QAnon accounts is visualized in Fig. 4a. We separately plot the total daily volume of tweets and the daily volume of tweets containing the QAnon keywords (frequently associated with QAnon listed earlier in Table 5). As observed, although the volume of tweets with QAnon associated keywords sharply declines after the Twitter Action, but the overall volume of activities from these accounts is sustained even after imposed bans and restrictions. This clearly suggests a gap in enforcement actions and desired outcomes. So we investigate evasion strategies that rendered Twitter Action ineffective.

In Fig. 4b we inspect whether new QAnon accounts were injected to continue activities after the Twitter Action. Whilst a declining and small fraction of new accounts were introduced, the fraction of the total volume of QAnon account tweets attributed to new accounts created after Twitter Action is less than 3%. Clearly, much of the volume is sustained by earlier accounts even after the ban.

**Evasion strategies with emerging and declining hashtags.** Twitter restrictions attempted to make QAnon content less visible, by banning QAnon-related content from appearing in trends and search. Therefore, we examine changes in QAnon content through analysis of hashtag usage patterns, that could have been employed as evasion strategies to sidestep Twitter restrictions. To that end, we leverage regression discontinuity design (RDD) to estimate causal effects of Twitter intervention, on hashtags adopted by QAnon accounts. RDD is a design which elicits causal effects of interventions by assigning a threshold above/below which an intervention is assigned \cite{Lee and Lemieux 2010}. By comparing observations lying closely on either side of the threshold, it is possible to estimate the average treatment effect from observed data, when randomized trials are infeasible.

### Table 8: Top 10 hashtags that declined ($\beta < 0$) in usage by QAnon post Twitter Action, estimated by regression discontinuity design in top 10K hashtags used by QAnon accounts

| Hashtag      | $|\beta|$  | slope | intercept | p-value |
|--------------|---------|-------|-----------|---------|
| wwg1wga      | 1.512   | -0.047| 4.951     | 0.002   |
| qanon        | 0.901   | -0.03 | 2.974     | 0.007   |
| kang         | 0.411   | -0.002| 1.149     | 0.000   |
| kag          | 0.226   | -0.004| 0.584     | 0.000   |
| q            | 0.207   | -0.008| 0.7       | 0.048   |
| qarmy        | 0.139   | -0.005| 0.47      | 0.071   |
| wwg1gwa      | 0.108   | -0.004| 0.42      | 0.031   |
| qanos        | 0.087   | -0.001| 0.164     | 0.010   |
| patriotstrike |0.082   | -0.001| 0.233     | 0.000   |
| inittogther  | 0.067   | -0.002| 0.19      | 0.083   |

For RDD, we consider each hashtag in the most frequent 10K hashtags adopted in QAnon account tweets, and fit a linear regression model on the hashtag’s daily usage volume during the data collection period, with an additional treatment variable assigned to the regression model to capture intervention effect. The regression model is as follows, $y = mx + b + \beta \mathbb{1}\{x > x_0\}$. Here, the hashtag usage volume per day (i.e., how many times the hashtag was adopted in tweets from QAnon accounts), normalized by tweet volume from QAnon accounts on that day, is modeled by dependent variable $y$, over each day $x$. The slope $m$ and intercept $b$ captures the trend in hashtag usage, and the coefficient $\beta$ of the treatment variable $\mathbb{1}\{x > x_0\}$ captures the discontinuity at threshold $x_0$, selected as the end of the week that Twitter announced and enforced its restrictions on QAnon.

In Table 8 and 9 we list identified hashtags with highest estimated treatment effects $|\beta|$ (i.e., most change with respect to Twitter’s intervention on QAnon with p-value of $\beta$ at most 0.2) and highlighted in bold are p-value $\leq 0.05$ i.e., 95% confidence interval.\footnote{The goodness of fit tests and other regression statistics, with few illustrated examples of identified hashtags are in the Appendix.}

Figure 4: QAnon activity and account creation timeline with respect to Twitter Action banning QAnon, DNC and RNC.
Table 9: Top 20 hashtags that increased ($\beta > 0$) in usage by QAnon post Twitter Action, estimated by regression discontinuity design in top 10K hashtags used by QAnon accounts

| Hashtag                  | $|\beta|$ | slope   | intercept | p-value |
|--------------------------|--------|---------|-----------|---------|
| walkawayfromdemocrats    | 0.138  | -0.001  | 0.145     | 0.000   |
| saveourchildren          | 0.099  | 0.002   | 0.011     | 0.158   |
| huge                     | 0.074  | -0.001  | 0.016     | 0.130   |
| heelsupharris             | 0.068  | 0.0     | 0.006     | 0.098   |
| vote                     | 0.067  | 0.0     | 0.071     | 0.001   |
| warroombandemic          | 0.065  | -0.001  | 0.041     | 0.007   |
| th3d3n                   | 0.064  | -0.001  | 0.016     | 0.002   |
| womenfortrump            | 0.061  | 0.0     | 0.086     | 0.133   |
| thesilentmajority        | 0.059  | 0.0     | 0.017     | 0.122   |
| sallyaytes               | 0.058  | -0.001  | 0.015     | 0.058   |
| hcqworksfauciknewin2005  | 0.055  | -0.001  | 0.014     | 0.000   |
| nomalinvoting            | 0.053  | 0.0     | 0.014     | 0.000   |
| mo03                     | 0.051  | -0.001  | 0.015     | 0.032   |
| trump2020victory         | 0.05   | -0.001  | 0.03      | 0.001   |
| hermancain               | 0.05   | -0.001  | 0.016     | 0.003   |
| bigharma                 | 0.05   | -0.001  | 0.018     | 0.000   |
| jimcrowjoe               | 0.016  | -0.001  | 0.032     | 0.072   |
| bidenisappedo            | 0.049  | -0.001  | 0.014     | 0.011   |
| vppick                   | 0.049  | -0.001  | 0.008     | 0.172   |
| hcqzinc4prevention       | 0.048  | -0.001  | 0.015     | 0.000   |

Table 10: % Decrease in ratio of direct engagements (reply to, retweet, quote) with QAnon accounts tweets to volume of QAnon accounts tweets, before and after Twitter Action

| Tweet Type | Vol./day (Before) | Vol./day (After) | % Decrease (Vol./day) | %Decrease (Eng vol./QAnon tweet) |
|------------|-------------------|------------------|----------------------|-------------------------------|
| QAnon      | 295,267           | 361,176          | 22.32                | -                             |
| RP.TO      | 49,161            | 46,114           | -6.2                 | 23.32                         |
| RT         | 177,013           | 189,387          | 6.99                 | 12.53                         |
| QTD        | 8908              | 8985             | 0.86                 | 17.55                         |

Table 10: % Decrease in ratio of direct engagements (reply to, retweet, quote) with QAnon accounts tweets to volume of QAnon accounts tweets, before and after Twitter Action.

Table 9: Top 20 hashtags that increased ($\beta > 0$) in usage by QAnon post Twitter Action, estimated by regression discontinuity design in top 10K hashtags used by QAnon accounts.

Table 10: % Decrease in ratio of direct engagements (reply to, retweet, quote) with QAnon accounts tweets to volume of QAnon accounts tweets, before and after Twitter Action.

Table 10: % Decrease in ratio of direct engagements (reply to, retweet, quote) with QAnon accounts tweets to volume of QAnon accounts tweets, before and after Twitter Action.

Table 9: Top 20 hashtags that increased ($\beta > 0$) in usage by QAnon post Twitter Action, estimated by regression discontinuity design in top 10K hashtags used by QAnon accounts.

Table 10: % Decrease in ratio of direct engagements (reply to, retweet, quote) with QAnon accounts tweets to volume of QAnon accounts tweets, before and after Twitter Action.

Table 9: Top 20 hashtags that increased ($\beta > 0$) in usage by QAnon post Twitter Action, estimated by regression discontinuity design in top 10K hashtags used by QAnon accounts.

Table 10: % Decrease in ratio of direct engagements (reply to, retweet, quote) with QAnon accounts tweets to volume of QAnon accounts tweets, before and after Twitter Action.

Table 9: Top 20 hashtags that increased ($\beta > 0$) in usage by QAnon post Twitter Action, estimated by regression discontinuity design in top 10K hashtags used by QAnon accounts.

Table 10: % Decrease in ratio of direct engagements (reply to, retweet, quote) with QAnon accounts tweets to volume of QAnon accounts tweets, before and after Twitter Action.

Table 9: Top 20 hashtags that increased ($\beta > 0$) in usage by QAnon post Twitter Action, estimated by regression discontinuity design in top 10K hashtags used by QAnon accounts.

Table 10: % Decrease in ratio of direct engagements (reply to, retweet, quote) with QAnon accounts tweets to volume of QAnon accounts tweets, before and after Twitter Action.
We also discuss inherent limitations in data collection. The standard API allows access to a ~1% sample of the Twitter stream filtered based on tracked parameters. Since Twitter does not allow unpaid access to historical tweets older than 7 days, it is hard to do iterative/snowball sampling based on keyword expansions. Therefore, we rely on tracking candidate mentions in tweet text/metadata instead of manually selected fixed keywords. The inspected topics, accounts, and hashtags distributions do suggest comprehensive coverage of election events and disinformation.

**Appendix**

**Regression statistics.** In Table 1, goodness-of-fit tests are provided for the RDD regression function (illustrated examples in Fig.6). When comparing different degree polynomials as chosen functions for RDD, we found AIC and BIC were lowest for degree-1 (we measured up to degree-4, since the results are similar we show only degree-1 and degree-2 in the table). Degree-1 is the model used in the RDD analysis $y = mx + b + \beta I(x > x_0)$ (Lee and Lemieux 2010). Adjusted $R^2$, f-statistic and p-values of the coefficients also indicate degree-1 is preferred, since the p-value of coefficient $m_2$ for $x^2$ term is not significant at 0.05. The results were also consistent when separate coefficients of dependent variables were used on either side of the intervention.

**Cascade propagation.** To confirm that our findings are not...
biased by CSI predictions, in Fig 7, we compare only labeled set of unreliable/conspiracy vs. reliable cascades (labeled by the news source, used in training CSI). We find the conclusions are robust with similar trends for all propagation properties. CSI uses engagements, but also text, temporal and account suspiciousness features, making predictions more robust, and verified to be within reasonable error rates.

References

Badawy, A.; Addawood, A.; Lerman, K.; and Ferrara, E. 2019. Characterizing the 2016 Russian IRA influence campaign. SWAM .

Bessi, A.; and Ferrara, E. 2016. Social bots distort the 2016 US Presidential election online discussion. First Monday 21(11).

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008(10): P10008.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching Word Vectors with Subword Information. TACL 5.

Chen, E.; Deb, A.; and Ferrara, E. 2020. Election2020: The First Public Twitter Dataset on the 2020 US Presidential Election.

Cresci, S. 2020. A decade of social bot detection. ACM 63(10): 72–83. ISSN 1557-7317.

de Zeeuw, D.; Hagen, S.; Peeters, S.; and Jokubauskaite, E. 2020. Tracing normiefication. First Monday.

Ferrara, E.; Chang, H.; Chen, E.; Muric, G.; and Patel, J. 2020. Characterizing social media manipulation in the 2020 US Presidential election. First Monday.

Friggeri, A.; Adamic, L. A.; Eckles, D.; and Cheng, J. 2014. Rumor Cascades. In ICWSM.