Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

The Manufacture of Partisan Echo Chambers by Follow Train Abuse on Twitter

Christopher Torres-Lugo, Kai-Cheng Yang, Filippo Menczer

Observatory on Social Media, Indiana University, Bloomington, USA
{torresch, yangkc, fil}@iu.edu

Abstract

A growing body of evidence points to critical vulnerabilities of social media, such as the emergence of partisan echo chambers and the viral spread of misinformation. We show that these vulnerabilities are amplified by abusive behaviors associated with so-called "political follow trains" on Twitter, in which long lists of ideologically-aligned accounts are mentioned for others to follow. We present the first systematic analysis of a large U.S. hyper-partisan train network. We observe an artificial inflation of influence: accounts heavily promoted by follow trains profit from a median six-fold increase in daily follower growth. This catalyzes the formation of highly clustered echo chambers, hierarchically organized around a dense core of active accounts. Train accounts also engage in other behaviors that violate platform policies: we find evidence of activity by inauthentic automated accounts and abnormal content deletion, as well as amplification of toxic content from low-credibility and conspiratorial sources. Some train accounts have been active for years, suggesting that platforms need to pay greater attention to this kind of abuse.

Introduction

In the past decade, online social media have become an important platform for participating in social movements regarding various public issues like economic inequality (Gleason 2013; Conover et al. 2013), human rights (Stewart et al. 2017), and especially political elections (Flores-Saviaga, Keegan, and Savage 2018). Features of social media like peer-to-peer communication, anonymity, high efficiency, and broad coverage greatly facilitate organization efforts (Lotan et al. 2011; Starbird and Palen 2012; Tufekci and Wilson 2012). Unfortunately, the same features invite problems like inauthentic behaviors (Ferrara et al. 2016; Pacheco et al. 2021), echo chambers (Sasahara et al. 2020), and the wide spread of toxic information (Shao et al. 2018; Grinberg et al. 2019; Ahmed et al. 2020) that challenge the integrity of the information ecosystem. The January 6, 2021 riot at the U.S. Capitol provides clear evidence of the entanglement between online disinformation and real-world harm,[1] making it all the more crucial to study how bad actors manipulate online discourse.

A Twitter account may occasionally recommend other accounts to their followers. This behavior, as flagged by the popular *#FF* (follow Friday) hashtag, is legitimate. "Follow trains," on the other hand, are a manipulative form of this behavior with the intent of amassing followers. Follow trains are defined by Twitter as a type of "reciprocal inflation — trading or coordinating to exchange follows" and explicitly prohibited by the platform's manipulation and spam policy.[2]

Our own observations suggest that follow trains are widely abused for political manipulation on Twitter and other social media platforms. The characteristic behavior of partisan follow trains is the publishing of spam-like "train tweets" that typically contain a list of mentions (accounts), a media object, and possibly a few words. The screenshots in Figure 1 show examples of two political follow train tweets. The organizers ("train conductors") post the train tweets so that their followers can follow the accounts being mentioned ("train riders").

Tweets from train accounts are used to constantly seek attention from government officials and politicians by means of mentioning, retweeting, and replying. Occasionally they get amplified by influential users and reach a much broader audience. For example, President Trump retweeted train accounts in the past.[3]

Toward the goal of understanding partisan follow trains, we present, to the best of our knowledge, the first systematic analysis of such abuse. We focus on the Twitter platform and specifically pro-Trump follow trains. We present an analytical characterization of train accounts from multiple perspectives, with an emphasis on their social networks and questionable behaviors. We build and share datasets of train accounts, collect their tweets, and contrast their behaviors with those of accounts in baseline datasets to provide meaningful contexts for our analysis. This paper contributes the following findings:

- Pro-Trump follow trains are highly effective in inflating follower numbers for the train riders.

[1] apnews.com/article/donald-trump-conspiracy-theories-michael-pence-media-social-media-daba3f5dd16a431abc627a5cfc922b87

[2] help.twitter.com/en/rules-and-policies/platform-manipulation

[3] See web.archive.org/web/20200410122951/https://twitter.com/realdonaldtrump/status/1248589007329595392 for an archived example (Trump's account is suspended as of this writing).

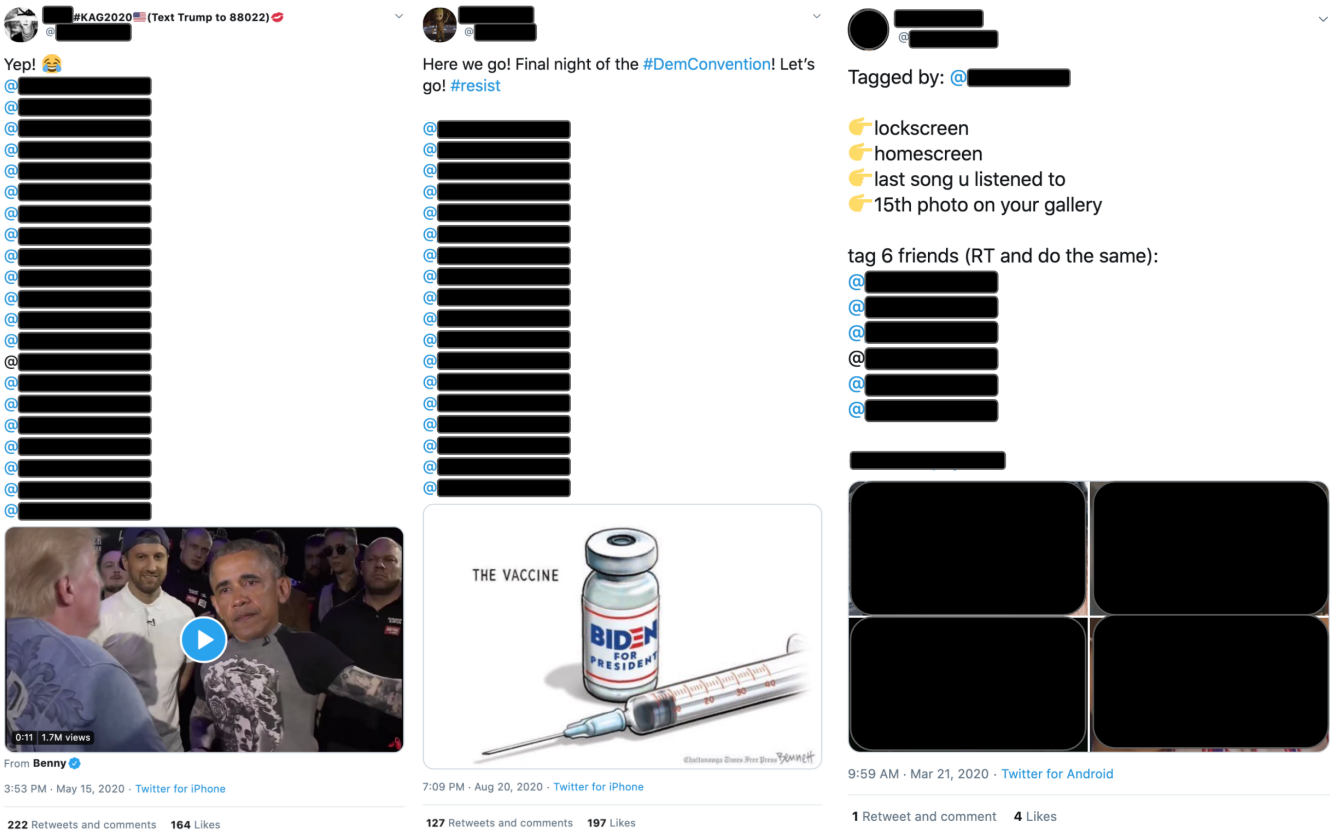

Figure 1: Screenshots of exemplar tweets from (left) a pro-Trump train conductor, (center) an anti-Trump train conductor, and (right) the tagging game. Some account information is redacted to protect privacy.

- Train accounts are more active than baseline pro-Trump Twitter users in terms of establishing social ties. As a result, they form a hierarchical and dense community that is highly coordinated, persistent, homogeneous, and fully focused on amplifying conservative narratives. These are characteristics that we use to label the community as a partisan echo chamber (Sasahara et al. 2020).
- Train accounts are also more active than baseline pro-Trump users in terms of posting tweets.
- In addition to generating and amplifying a large amount of spam-like tweets, some of the accounts abuse the platform through inauthentic personas and abnormal tweet deletions.
- By analyzing their tweets, we find that train accounts actively share a large volume of toxic information, such as low-credibility news and conspiracy theories.
- Finally, we observe the same behavioral patterns when replicating the analysis on an anti-Trump follow train network, suggesting the findings can generalize to other political networks.

Background

Social media influence stems from having high visibility on a platform. Despite early analysis showing that influence is not exclusively determined by the number of followers (Cha et al. 2010), it is generally assumed that users need to have many followers to increase their visibility. Some unscrupulous actors therefore resort to follower growth hacking. A well-studied growth hack is to purchase "fake followers," which often consist of inauthentic or compromised accounts (Cresci et al. 2015; Aggarwal and Kumaraguru 2015). There are also reports of organized exchanges to turn unpaid customers and volunteers into fake followers (Stringhini et al. 2013; Liu et al. 2016). Follow trains are a particularly effective follower growth hack: actions are coordinated with the ultimate goal of building a well-connected community with maximal influence.

In addition to violating Twitter's policies, partisan follow trains may also produce undesirable outcomes. For example, polarized and segregated echo chambers are commonly observed on social media (Jamieson and Cappella 2008; Garrett 2009; Conover et al. 2011; Lee et al. 2014), possibly leading to radicalization (Wojcieszak 2010; Bright 2018). Behind the curtain is the interplay between social biases like the tendency to establish belief-consistent social ties (Del Vicario et al. 2017; Hills 2019), cognitive biases such as information overload and confirmation bias (Menczer and Hills 2020), and social media mechanisms like friend recommendations and the ease of

(un)following (Sasahara et al. 2020). Blindly following riders recommended by like-minded conductors can easily accelerate the formation of polarized echo chambers.

Partisan follow trains also make the online community more vulnerable to inauthentic actors, who are blindly followed when recommended. Well-known types of inauthentic accounts include trolls (Zannettou et al. 2019) and malicious social bots (Ferrara et al. 2016), which have been actively involved in online discussions of elections across various countries (Bessi and Ferrara 2016; Deb et al. 2019; Stella, Ferrara, and De Domenico 2018; Ferrara 2017; Badawy, Ferrara, and Lerman 2018; Zannettou et al. 2019). Follow trains provide an easy mechanism for an entity to control automated accounts programmed to follow accounts mentioned by a conductor. Even train conductors can be automated. Recently, more attention has been drawn toward a new type of inauthentic actors that act in a coordinated fashion to increase influence and evade detection (Nizzoli et al. 2021; Sharma et al. 2021; Pacheco et al. 2021). Follow trains facilitate the formation of coordinated inauthentic networks.

Another problem regarding follow trains is the spread of toxic information, such as conspiracy theories and misinformation. Concerns about "fake news" on social media have been growing since the 2016 U.S. presidential election (Lazer et al. 2018; Vosoughi, Roy, and Aral 2018; Grinberg et al. 2019; Bovet and Makse 2019). During the 2020 COVID-19 pandemic, misinformation related to the outbreak, also known as the "infodemic," has also spread virally (Zarocostas 2020; Yang, Torres-Lugo, and Menczer 2020; Yang et al. 2021). Recent studies show that polarized echo chambers are associated with the diffusion of misinformation (Del Vicario et al. 2016) and inauthentic actors like malicious bots are responsible for spreading low-credibility information related to politics (Shao et al. 2018); this suggests that partisan follow trains may also exacerbate the misinformation problem — a conjecture we explore in this paper. Due to the potential real-world consequences of misinformation and conspiracy theories about topics such as health and elections, it's important to thoroughly investigate the role of partisan follow train in such abuse.

Data Collection

All data and code necessary to reproduce the results in this paper are shared in a public repository.[4]

Network Analysis

First, we focus on the social (mention and follow) networks of train accounts. We utilize snowball sampling to crawl a mention network of partisan train accounts. Since the focus of the present paper is on pro-Trump follow trains, we start the crawl from a high-profile pro-Trump train account. We query the Twitter search API to retrieve this account's tweets (retweets and replies are excluded) that contain user mentions. To ensure convergence of the snowball sampling and to focus on the most influential and suspicious accounts, we extract accounts that were mentioned in a tweet that contains at least nine mentions and was retweeted at least 40 times. The procedure is repeated recursively on the newly extracted accounts, until no new accounts emerge.

The two thresholds for the minimum numbers of mentions and retweets were manually selected after a close examination of multiple high-profile train accounts. We conducted a *post-facto* analysis of the distribution of the number of users mentioned in train tweets sampled from a historical collection. The distribution is bimodal and the selected threshold (nine or more mentions) lies between the two modes. Therefore the threshold allows us to separate two groups: *train conductors* who engage in mentioning a large number of accounts and *train riders* (the rest).

To further refine the dataset, we conduct an exhaustive manual annotation of the accounts collected. We find that 3% are suspended or deleted and 1% are non-political or inactive. These accounts are removed because we couldn't obtain enough data about them to perform further analysis. In addition, 9% of the accounts are verified; these are also excluded from further analysis because they include celebrities whose influence would bias our analysis on numbers and growth of followers. The remaining accounts are hyper-partisan. Among them we find a small group of anti-Trump accounts. Further inspection reveals that this is due to a pro-Trump account mentioning an anti-Trump account. The anti-Trump portion of the network is also excluded from the present analysis.

The data collection took place between February 16 and 26, 2020. The resulting mention network, denoted as `train-net`, contains 8,308 nodes (182 conductors and 8,126 riders) and 20,773 edges. Note that this categorization only reflects the behaviors of the accounts in the data collection period; riders could act as conductors at a different time. Figure 2 shows common hashtags in the profile descriptions of the train accounts in the network, demonstrating a clear alignment with pro-Trump themes.

To gauge the structure of the mention network, we need a suitable baseline. We use an analogous method to collect a group of accounts that share similar mention behaviors but in a non-political context. We take advantage of an online game played by many Twitter users in the early days of the 2020 COVID-19 lockdown. In the game, each tagged user is asked to post a screenshot of their phone and mention six friends to continue the game with the same instruction. An exemplar tagging tweet can be seen in Figure 1. We identify game participants through phrases like *"tagged by," "lockscreen," "homescreen,"* and *"last song u listened to"*; other collected accounts are removed. The data collection took place between March 21 and 27, 2020, roughly a month after the collection of `train-net`. Since the tagging game is not focused on any particular topics, we believe the time difference does not have a substantial impact on the analysis. The resulting mention network has 5,567 nodes and 7,189 edges. We denote this dataset as `tagging-net`.

Behavioral Analysis

The small number of conductor accounts in the `train-net` dataset makes it difficult to obtain a robust statistical analysis of their profiles and behaviors. To expand the number of conductors, we leverage the

[4] github.com/osome-iu/twitter-follow-trains

Figure 2: Frequent hashtags in account descriptions in (top) `train-net` and (bottom) pro-Trump baseline.

observation that accounts engaged in train behavior amplify follow train tweets by other conductors. We selected all the conductors and a similarly-sized random sample (18%) of the riders in `train-net`. The user IDs of these accounts were used to retrieve tweets posted by them between January 1, 2018 and June 30, 2020. This was done by querying a historical archive of the so-called Decahose, a 10% random sample of the public tweet stream (Davis et al. 2016). We scanned the resulting sample of tweets to extract retweets and quotes posted by the selected train accounts. From the retweeted or quoted tweets, we then selected those meeting our heuristic rules, yielding an additional set of train tweets. The accounts that published these train tweets were considered to be additional conductors, excluding verified accounts. We finally retrieved a sample of tweets by the additional conductors as well, in the same period (Jan 2018–Jun 2020), again using the historical archive.

While the train conductors and riders in our data are all pro-Trump, not all pro-Trump accounts are involved in follow trains. We therefore need another suitable baseline to examine whether the characteristics of the train accounts are typical of pro-Trump users. We build a representative sample of pro-Trump Twitter accounts starting from a list of hashtags frequently found in pro-Trump account descriptions from `train-net` (see Figure 2). We scan the historical archive and find all the accounts that tweeted on February 20, 2020 and that had any of the selected hashtags in their descriptions. Among the matched partisan accounts, we further sample users to approximately match the number of train accounts. Any sampled account that is part of the trains dataset is excluded. Figure 2 shows the most frequent hashtags in the descriptions of these baseline accounts. A clear partisan alignment with `train-net` accounts can be observed.

As for the pro-Trump train accounts, we collect historical

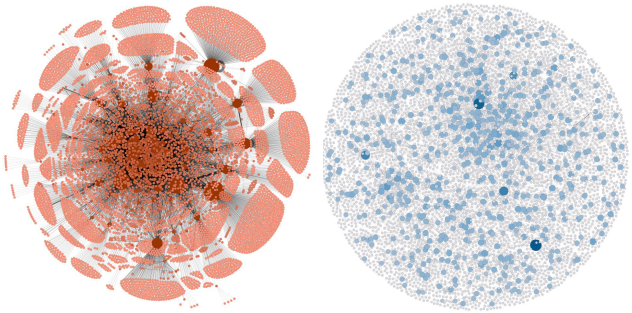

Figure 3: Visualizations of the mention networks among accounts in the (left) `train-net` dataset and (right) `tagging-net` baseline. We used the Fruchterman-Reingold layout algorithm (Fruchterman and Reingold 1991) in Gephi with identical parameters for both networks. Larger size and darker color of a node denote higher degree. In the train network, singleton nodes (riders mentioned by suspended conductor accounts) are not shown.

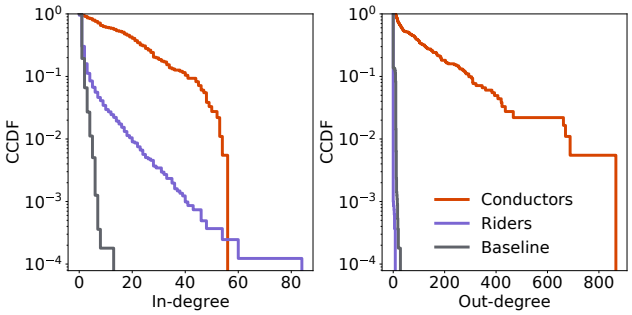

Figure 4: Complementary cumulative distributions of (left) in-degree and (right) out-degree for accounts in the `train-net` mention network and `tagging-net` baseline. Kolmogorov–Smirnov (KS) tests show that all of the distributions are significantly different from each other ($p < 0.01$).

tweets by the pro-Trump baseline accounts between January 1, 2018 and June 30, 2020. We denote the historical dataset containing all the conductors (those from `train-net` plus the additional ones), the sampled riders, and the pro-Trump baseline as `decahose`. It contains 14,693,650 tweets by 4,581 pro-Trump accounts: 1,342 conductors, 1,352 riders, and 1,839 baseline accounts.

Anatomy of Follow Train Networks

Mention Networks

Let us first analyze the `train-net` and `tagging-net` mention networks, in which each node represents an account and an edge from node A to B means that account A has mentioned B. Edge weights represent the numbers of mentions. Figure 3 visualizes the networks, showing a much more hierarchical structure in the follow train network compared to the baseline.

To characterize the mention network, the in-degree and

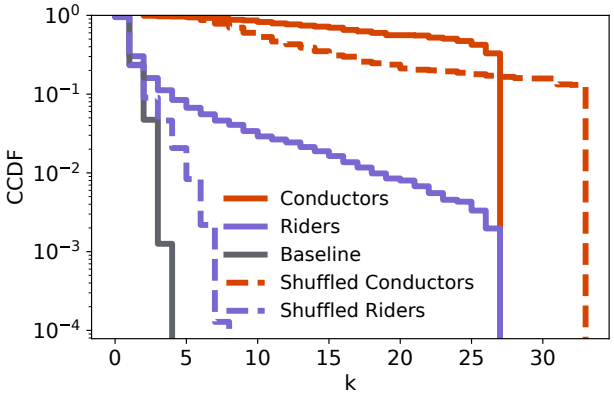

Figure 5: Complementary cumulative distribution of core number k for accounts in the `train-net` network and `tagging-net` baseline, along with a shuffled version of the `train-net` network. KS tests show that all distributions are significantly different from each other ($p < 0.01$).

out-degree distributions of the nodes are plotted in Figure 4. High in-degree indicates that an account was mentioned by many others and high out-degree means the account mentioned many others. We find that conductors tend to have much higher out-degree than rider and baseline accounts, consistent with their role. As a result, riders are mentioned more than baseline accounts.

We observe that the pro-Trump community in Figure 3 has a densely connected core and many peripheral nodes, while the `tagging-net` network displays a more homogeneous structure. To confirm this observation, we perform core decomposition of undirected versions of both networks and calculate the core number of each node, which measures node centrality and influence (Kitsak et al. 2010). Figure 5 plots the distributions of core number k. For the `tagging-net` baseline network, all nodes have k values smaller than five. The `train-net` network, on the contrary, has a deeply hierarchical structure with a very dense core (high k). Conductors tend to have higher core values than riders, indicating that they tend to be situated near the core of the network.

Nodes with high degree tend to have high core values. To disentangle the roles of degree and k values, we shuffle the edges of the `train-net` network in Figure 3 while preserving the degrees of the nodes and perform core decomposition again on the shuffled network. The core number distributions after shuffling are also shown in Figure 5. For the conductors, we observe that high core numbers after shuffling are even larger — in fact, we can see some high-degree nodes near the periphery of the network in Figure 3. This suggests that their position near the core of the network is consistent with their high mentioning activity. For riders, on the other hand, high core numbers are not explained by degree, suggesting that these accounts are pulled toward the core of the network by the conductors.

Let us examine the clustering structure of the networks.

	Conductors	Riders	Baseline	Anti-Trump
Conductors	50.3%	28.0%	17.0%	0%
Riders		8.2%	3.8%	0%
Baseline			2.8%	0%
Anti-Trump				2.0%

Table 1: Percentages of account pairs with follow edges within and across groups. "Conductors," "Riders," and "Baseline" refer to pro-Trump accounts sampled from the `decahose` dataset, while "anti-Trump" accounts are sampled using anti-Trump hashtags (see text).

While the `train-net` and `tagging-net` networks have similar densities (3.0×10^{-4} vs. 2.3×10^{-4}), the former has a much higher average clustering coefficient (0.13 vs. 0.03). The high number of triangles in the train network suggests that some rider accounts mention other riders, acting also as conductors.

In summary, the above analyses show that compared to the baseline, the mention network of pro-Trump trains is heavily clustered and hierarchically organized around a dense core of highly active conductors.

Follow Network

We are also interested in the follow network induced by the pro-Trump train accounts. Since querying the "follow" relationship between each pair of accounts is not practical due to Twitter API's rate limit, we adopt a sampling strategy. We split the pro-Trump accounts in `decahose` into conductor, rider, and baseline groups to examine the follow relations within and across groups.

Studies of political echo chambers have focused on the segregation between conservative and liberal communities on Twitter (Conover et al. 2011). Here we define an echo chamber as a network community of politically homogeneous accounts that are significantly more likely to be connected with each other than with politically discordant accounts. To explore this echo-chamber phenomenon, we consider an additional sample of anti-Trump accounts as a fourth group. We follow an analogous procedure to the baseline accounts in `decahose`. However, instead of using hashtags frequently found in descriptions of pro-Trump accounts, we query the historical archive using hashtags frequently found in descriptions of anti-Trump accounts encountered during our snowball crawl (see Data Collection Section).

From each of the four account groups, we sample 5,000 account pairs and use the Twitter friendship API to check whether the accounts in each pair are following each other. An account pair is considered to have a follow edge if either account follows the other.

We report the percentages of account pairs with follow edges within and across groups in Table 1. The anti- and pro-Trump (baseline) groups are clearly segregated, with no cross-connections. What is more interesting is the presence of denser echo chambers within the pro-Trump community, driven by train networks. In particular, riders are more likely to connect to each other than to baseline pro-Trump ac-

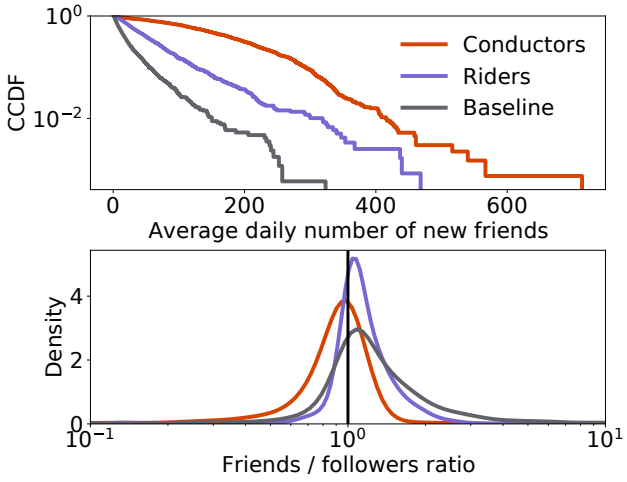

Figure 6: Top: Complementary cumulative distributions of average daily number of new friends for accounts in the decahose dataset. We excluded a few cases with a daily growth above the maximum allowed by Twitter, which may be due to API data errors. Bottom: Kernel Density Estimations (KDE) of the distributions of the friend/follower ratios. The vertical black line indicates the ratio of one. All distributions are significantly different (KS tests, $p < 0.01$).

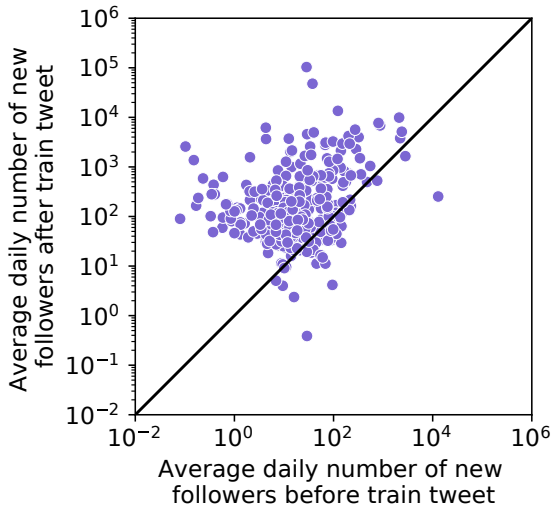

Figure 7: Daily number of new followers of rider accounts. The scatter plot compares the growth before (x-axis) and after (y-axis) the train tweet mentioning each rider. The numbers are obtained from rider tweets before and after a train tweet (see text). The diagonal shows the expected daily growth.

counts. And they are even more likely to connect to conductors. Finally, the conductors are more densely connected to baseline and rider accounts, and have highest likelihood to follow other conductors. These results confirm the role of follow trains in boosting the clustered structure of pro-Trump echo chambers.

Profile and Behavioral Characterization

Follow Manipulation

Let us examine the follow behavior of train accounts. In the decahose dataset, each tweet contains profile information that reflects the status of the author account at the time the tweet was posted. A series of tweets by the same account provide snapshots capturing the temporal evolution of the account's profile. By examining the difference in friend counts between tweets in consecutive days, we can estimate the account's daily growth in the number of friends. We plot the distributions of the average daily number of new friends for different groups of accounts in Figure 6 (top). Train accounts establish social ties much more aggressively than the baseline.

When an account has more than 5,000 friends, Twitter imposes constraints on their friend/follower ratio to prevent manipulation of follow relationships.[5] Although the exact ratio threshold is not published, it is generally understood that an account must have a friend/follower ratio below a critical value close to one. In other words, one can follow additional accounts only after having a similar number of followers. The mutual following patterns promoted by partisan trains are likely designed to circumvent this constraint: riders accumulate followers in order to aggressively follow new accounts. In fact, Figure 6 (bottom) shows that rider accounts have a friend/follower ratio that is narrowly distributed around one.

Is follow manipulation by partisan trains effective? Let us quantify the amplification in the number of new followers gained by rider accounts through train tweets. We focus on riders that (i) were mentioned by many train tweets in decahose, (ii) are still active, and (iii) tweeted within 24 hours before and 48 hours after being mentioned. These requirements are implemented to ensure that we could obtain the necessary information to perform the analysis. For each of these riders, we consider only one of the train tweets mentioning it. Let t_{train} be the timestamp of the train tweet. We identify two tweets by the rider. The first occurs at time t_{before} most immediately preceding the train tweet, i.e., minimizing $t_{\text{train}} - t_{\text{before}}$ s.t. $t_{\text{train}} - 24h < t_{\text{before}} < t_{\text{train}}$. Let f_{before} be the follower count extracted from this tweet. The second rider's tweet occurs at time t_{after} after the train tweet and is selected by maximizing the follower count, i.e., $t_{\text{after}} = \arg\max_t(f_t \text{ s.t. } t_{\text{train}} < t < t_{\text{train}} + 48h)$. The difference between the follower counts in these two tweets, divided by the time elapsed since the train tweet, is used to estimate the daily number of new followers gained by the rider after its mention in the train tweet:

$$\Delta_{\text{after}} = \frac{f_{\text{after}} - f_{\text{before}}}{t_{\text{after}} - t_{\text{train}}}.$$

This number is compared with the estimated daily number

[5] help.twitter.com/en/using-twitter/twitter-follow-limit

of new followers gained by the rider before the train tweet:

$$\Delta_{\text{before}} = \frac{f_{\text{before}}}{t_{\text{before}} - t_0},$$

where t_0 is the rider's creation timestamp.

Figure 7 plots Δ_{after} versus Δ_{before} for 394 riders meeting the above criteria. Accounts situated on the diagonal line have no change in follower growth. Most of the riders are above the diagonal line (significant per Mann-Whitney U test, $p < 0.01$), suggesting that they profit from an increased growth in followers after the train tweet mention. The ratio $\Delta_{\text{after}}/\Delta_{\text{before}}$ indicates a median amplification of the follower gain over 600%.

Accounts Suspension

Since the 2018 U.S. midterm election season, Twitter implemented more aggressive enforcement actions against policy violations and suspended accounts at a higher rate.[6] As the accounts involved in follow trains, especially the conductors, are violating platform policies, we are interested in their suspension rates. We checked the profile status of all accounts in `train-net` and `tagging-net` through the Twitter API on January 6, 2021. (Note that the `decahose` dataset is not suitable for this analysis because suspended and deleted accounts are removed from the historical archive.)

Since the original data collection in February 2020, 1,453 train accounts (15%) had been suspended and 582 (6%) deleted. Breaking these numbers by account type, 22.6% of the conductors and 14.8% of the riders were suspended; 7.7% of the conductors and 6% of the riders were deleted. By comparison, 1,334 (24%) of the accounts in the `tagging-net` baseline were suspended and 529 (9.5%) deleted since data collection in March 2020. These numbers suggest that, until January 6, 2021, train accounts had similar chances to be suspended as the non-political baseline. Twitter may have suspended more train accounts after the January 6 riots.[7]

Abusive Behaviors

In this section, we turn our focus to certain automated and abnormal behaviors of train accounts that may flag abuse.

Automation

As discussed in the Background section, various inauthentic actors might be involved in partisan follow trains; we focus on social bots here. To estimate the prevalence of automation, we adopt BotometerLite,[8] a scalable off-the-shelf bot detection tool (Yang et al. 2020). For each account, BotometerLite generates a bot score between 0 and 1 with higher values indicating more bot-like behaviors.

The bot score distributions in Figure 8 show that most baseline accounts in the `decahose` dataset are human-like.

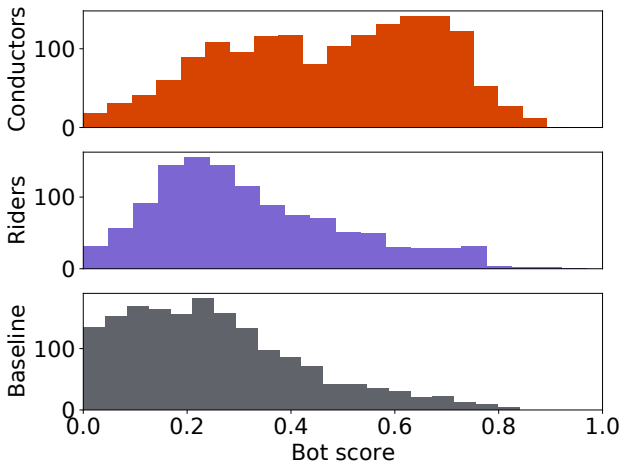

Figure 8: Bot score distributions of (top) conductor, (middle) rider, and (bottom) baseline accounts from the `decahose` dataset. One can use 0.5 as a threshold to binarize the classification results. Based on this threshold, the percentage of bot-like accounts is 39.4% for conductors, 12.0% for riders, and 5.3% for the baseline. The distributions are significantly different (KS tests, $p < 0.01$).

However, we do observe a larger number of bot-like accounts among train accounts and especially conductors, suggesting that follow trains may be sustained in part by social bots. Note that the action of posting train tweets alone does not have a substantial impact on the bot classification model.

Abnormal Deletion Behaviors

An examination of the user timelines of some of the train accounts shows that they publish a significant volume of tweets. In addition to this, we noticed that some of the train accounts routinely delete their tweets in bulk. Users have the freedom to delete their posts and may have legitimate reasons to do so (Almuhimedi et al. 2013). However, the deletion feature can also be abused. For example, trolls and malicious bots may delete their tweets to conceal their activities and intentions and evade detection (Zannettou et al. 2019; Yang et al. 2019; Elmas et al. 2019).

Let us use the `decahose` dataset to quantify the tweet publishing and deletion events. Each tweet contains an associated user object with information that reflects the status of the account when the tweet was posted. Although the data only contains samples of the tweets from each account, these can still provide multiple snapshots of an account at different times. A decrease or increase in the tweet count between snapshots of an account in consecutive days indicates tweet deletion or new published tweets, respectively. Note that the estimates obtained through this method provide a *lower bound* on the number of new tweets and deletions; the true numbers could be much larger, as a user could post and delete many tweets between consecutive snapshots.

Figure 9 shows the distributions of the estimated numbers of daily tweets deleted and published by accounts in the `decahose` dataset. Conductor accounts perform tweet

[6] transparency.twitter.com/en/reports/rules-enforcement.html

[7] blog.twitter.com/en_us/topics/company/2021/protecting--the-conversation-following-the-riots-in-washington--.html

[8] botometer.osome.iu.edu/botometerlite

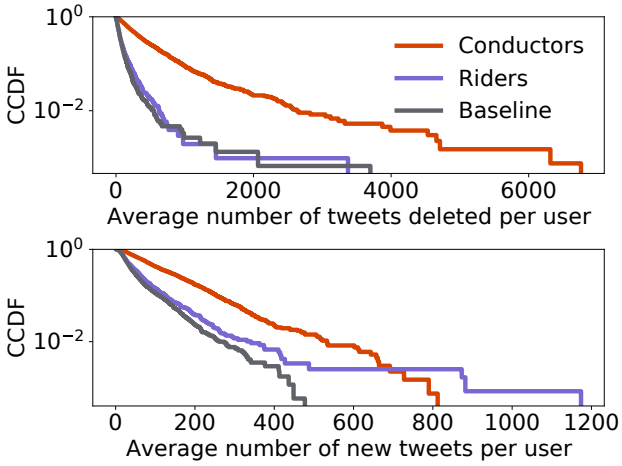

Figure 9: Complementary cumulative distribution of (top) average estimated daily tweet deletions per account and (bottom) average estimated daily new tweets per account in the decahose dataset. The distributions are significantly different (KS tests, $p < 0.01$), except between riders and baseline on tweet deletions.

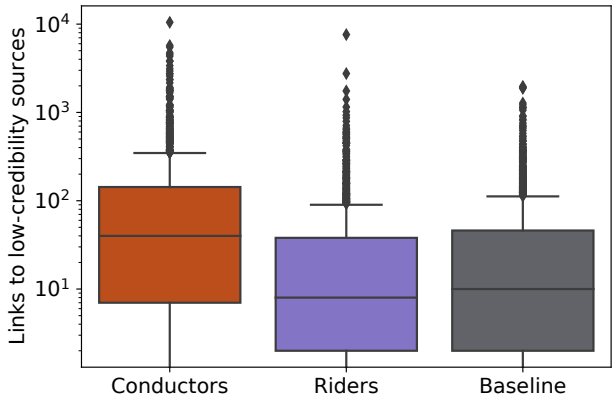

Figure 10: Boxplots showing the distributions of the numbers of links to low-credibility sources shared by accounts in decahose. Conductors share significantly more low-quality links than rider or baseline accounts (U tests, $p < 0.01$), whereas the rider and baseline distributions are not significantly different from each other.

deletion at a staggering frequency: on average, they delete at least 420 tweets per day. In contrast, Almuhimedi et al. (2013) show that typical Twitter users delete less than two tweets per day on average. The tweet deletion rates of conductor accounts are extremely high even in comparison to rider and baseline accounts (83 and 73 deleted tweets per day on average, respectively). And the tails of the distributions highlight accounts with thousands of deleted tweets per day — far exceeding the maximum number of 2,400 posts per day allowed by the platform.[9] Although Twitter terms forbid inspection of the deleted content,[10] such abnormal behaviors are strongly suggestive of abuse. Train accounts also produce higher volumes of tweets compared to the baseline.

Spreading Toxic Information

In this section, we analyze the spread of low-credibility news and conspiracy theories by partisan train accounts.

Low-credibility News

We identify low-credibility news based on sources. This approach is widely adopted in the literature (Shao et al. 2018; Grinberg et al. 2019; Pennycook and Rand 2019; Bovet and Makse 2019; Yang, Torres-Lugo, and Menczer 2020) because labeling at the level of individual articles is not feasible (Lazer et al. 2018). We use the *Iffy+* list of low-credibility sources.[11] *Iffy+* merges lists of sites that regularly publish mis/disinformation, as identified by major fact-checking and journalism organizations such as Media Bias/Fact Check, FactCheck.org, PolitiFact, BuzzFeed News, and Wikipedia.

We find links to low-credibility news outlets embedded in the decahose tweets. For each account, we count the *total* number of links to low-credibility sources per user. Figure 10 shows that conductors tend to share more of these links compared to rider and baseline accounts. Normalizing by account age yields similar results for *daily* numbers of low-quality link shares.

Conspiracy Theories

Some conspiracy theories have gained significant attention in the context of the 2020 U.S. presidential election.[12] A particularly notorious and dangerous conspiracy theory is QAnon, which was once only accepted by fringe groups. As we write this paper, multiple media have reported how QAnon has become more mainstream, merged with false narratives about the pandemic and the election, led to violence, and started to affect people's lives.[13] Popular social media platforms, including Facebook[14] and Twitter,[15] recently banned QAnon accounts, pages, and groups.

The frequent hashtags in Figure 2 suggest that some of the partisan train accounts label themselves as QAnon believers. To quantify the involvement of train accounts with QAnon content, we manually curate a list of QAnon keywords (the full list and details are available in the code and data repository).

[9] help.twitter.com/en/rules-and-policies/twitter-limits

[10] developer.twitter.com/en/docs/twitter-api/enterprise/compliance-firehose-api/overview

[11] iffy.news/iffy-plus

[12] osome.iu.edu/research/survey/files/FinalSummary_UnsupportedNarratives_OSoMe_a.pdf

[13] poynter.org/fact-checking/2021/a-man-wearing-a-buffalo-cap-proves-how-far-mis-disinformation-can-go-and-how-dangerous-it-can-be

[14] about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence

[15] twitter.com/TwitterSafety/status/1285726277719199746

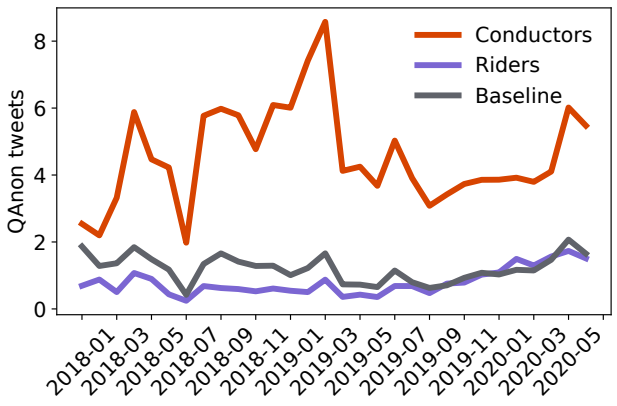

Figure 11: Monthly average numbers of `decahose` tweets per account containing QAnon-related keywords. The plot only considers the accounts created until each point in time.

To check the rate at which QAnon-related content is generated, the keywords are matched against the tweets in the `decahose` dataset. We then calculate the monthly average number of QAnon-related tweets for each group and show its time evolution in Figure 11. We observe that conductor accounts produce QAnon content at higher rates than rider and baseline accounts.

Generalizability of the Findings

The present paper focuses on the pro-Trump follow train network. Although the comparison with the tagging game network suggests that not all follow networks demonstrate the same characteristics, it is still interesting to test whether the findings generalize to other political follow networks. To this end, we analyze a smaller follow train network of anti-Trump accounts collected using the same methods described in the Data Collection section. The only difference is that we start the crawling with an anti-Trump account this time. We replicate the analyses and find that the results about network structure, follow behavior, and automated activity are consistent with those of the pro-Trump network (see Appendix).

Discussion

This paper provides an in-depth analysis of accounts involved in U.S. partisan follow trains on Twitter. Learning the characteristics of such accounts can help platforms, researchers, policymakers, and the general public recognize follow train abuse and take appropriate measures to curb its harm.

To the best of our knowledge, this is the first systematic investigation of political follow trains. Prior studies have focused on other follower growth hacks like purchasing fake followers. In contrast to these hacks, political follow trains manufacture a dense, clustered, and hierarchical echo chamber organized around a small core of conductor accounts. The influence of accounts in this echo chamber is boosted by manipulating follower relationships in ways that circumvent platform rules. Political train accounts are more likely to display inauthentic and abusive behaviors, such as high-volume posting and deletions, compared to ordinary partisan accounts with similar descriptions. They are also responsible for spreading low-credibility news as well as conspiracy theories. Part of this difference could be attributed to conductors being more active accounts, which could skew the sample from the historical archive to include more of their tweets.

Such abusive behaviors negatively affect the online experience of ordinary social media users who are exposed to false and inflammatory information. The real-world consequences are clearly demonstrated by the January 2021 attack on the U.S. Capitol, fueled by a systematic spread of election disinformation, such as QAnon conspiracies amplified by train accounts.

The partisan train phenomenon poses new challenges to social media platforms. Moderation is needed to mitigate the undesirable outcomes of partisan trains. Although Twitter has stepped up their efforts to maintain a healthy online discussion around critical issues like elections and public health, our findings suggest that at the time of our analysis, aggressive actions had not yet been taken to target this particular type of abusive behavior. For example, Twitter mentions in their Following FAQ[16] that inauthentic follows by third-party apps can result in account suspension. However, the same behavior by conductor accounts — whether perpetrated by automated apps or manually — did not lead to the prompt suspension of rider accounts. We believe that moderation policies could be broadened to target abusive follow train strategies.

Although the present study focuses on the pro-Trump follow train network on Twitter, similar abusive behaviors exist in other political follow trains as well, suggested by our analyses on an anti-Trump train network. Follow trains can also be found on other platforms like Facebook and Instagram, in other countries, and different languages. Our framework could be applied to extend the present analysis to different platforms and contexts, provided that data from such platforms is available.

Appendix: Anti-Trump Trains Figures

To test whether the findings from the pro-Trump follow trains generalize to other political follow networks, let us consider a smaller follow train network of anti-Trump accounts and its corresponding decahose baseline. The collections are based on the same methods described in the Data Collection section, except that the crawler starts from an anti-Trump account. The results about profile descriptions (Figure 12), network structure (Figures 13 and 14), follow behavior (Figure 15), and automated activity (Figure 16) are consistent with those of the pro-Trump network.

[16] help.twitter.com/en/using-twitter/following-faqs

Figure 12: Frequent hashtags in anti-Trump account descriptions in (top) follow train network and (bottom) decahose baseline.

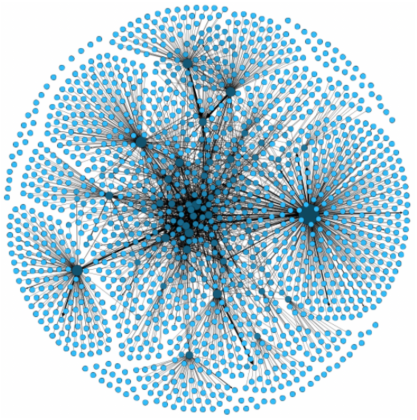

Figure 13: Visualization of the anti-Trump mention network. It has 1,545 nodes and 2,436 edges. Of the nodes, 50 are conductors and 1,495 are riders.

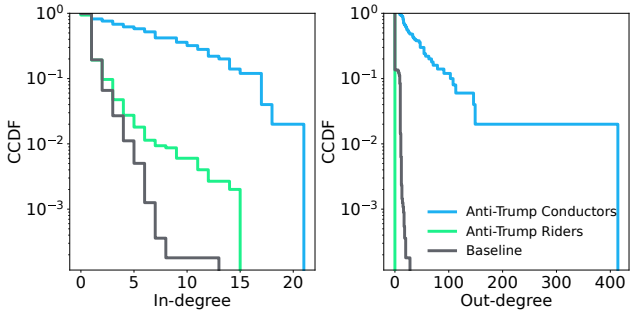

Figure 14: Complementary cumulative distributions of (left) in-degree and (right) out-degree for accounts in the anti-Trump train mention network. The results from the `tagging-net` baseline are also shown for comparison.

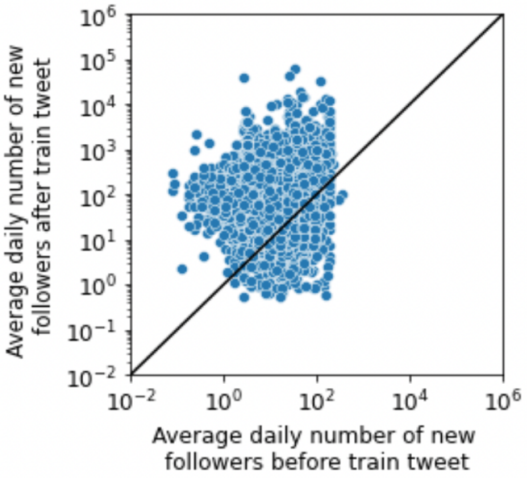

Figure 15: Daily number of new followers of rider accounts in the anti-Trump follow train network. The scatter plot compares the growth before (x-axis) and after (y-axis) the train tweet mentioning each rider. The numbers are obtained from rider tweets before and after a train tweet using the same procedure described in the main text. The diagonal shows the expected daily growth.

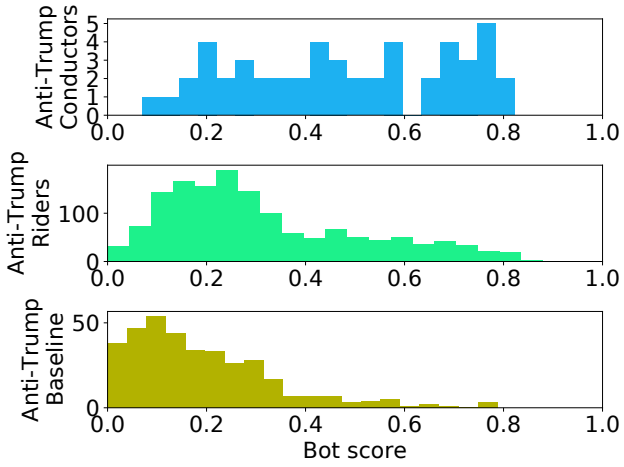

Figure 16: Bot score distributions of (top) conductor, (middle) rider, and (bottom) baseline accounts for anti-Trump accounts. By categorizing accounts with bot scores above 0.5 as bots, the percentage of bot-like accounts is 44.0% for conductors, 19.0% for riders, and 4.7% for the baseline. The distributions are significantly different (KS tests, $p < 0.01$).

Acknowledgments

This work was supported in part by Knight Foundation and Craig Newmark Philanthropies. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Aggarwal, A.; and Kumaraguru, P. 2015. What they do in shadows: Twitter underground follower market. In *Annual International on Privacy, Security and Trust*, 93–100.

Ahmed, W.; Vidal-Alaball, J.; Downing, J.; and Seguí, F. L. 2020. COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data. *Journal of Medical Internet Research*, 22(5): e19458.

Almuhimedi, H.; Wilson, S.; Liu, B.; Sadeh, N.; and Acquisti, A. 2013. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proceedings of the 2013 International Conference on Computer Supported Cooperative Work*, 897–908.

Badawy, A.; Ferrara, E.; and 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*, 258–265.

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

Bovet, A.; and Makse, H. A. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, 10(1): 1–14.

Bright, J. 2018. Explaining the Emergence of Political Fragmentation on Social Media: The Role of Ideology and Extremism. *Journal of Computer-Mediated Communication*, 23(1): 17–33.

Cha, M.; Haddadi, H.; Benevenuto, F.; Gummadi, P. K.; et al. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the Fourth International AAAI International on Weblogs and Social Media*.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Flammini, A.; and Menczer, F. 2011. Political Polarization on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Conover, M. D.; Ferrara, E.; Menczer, F.; and Flammini, A. 2013. The digital evolution of occupy wall street. *PloS one*, 8(5): e64679.

Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; and Tesconi, M. 2015. Fame for sale: efficient detection of fake Twitter followers. *Decision Support Systems*, 80: 56–71.

Davis, C. A.; Ciampaglia, G. L.; Aiello, L. M.; Chung, K.; Conover, M. D.; Ferrara, E.; Flammini, A.; Fox, G. C.; Gao, X.; Gonçalves, B.; Grabowicz, P. A.; Hong, K.; Hui, P.; McCaulay, S.; McKelvey, K.; Meiss, M. R.; Patil, S.; Peli Kankanamalage, C.; Pentchev, V.; Qiu, J.; Ratkiewicz, J.; Rudnick, A.; Serrette, B.; Shiralkar, P.; Varol, O.; Weng, L.; Wu, T.; Younge, A. J.; and Menczer, F. 2016. OSoMe: The IUNI Observatory on Social Media. *PeerJ Computer Science*, 2: e87.

Deb, A.; Luceri, L.; Badawy, A.; and Ferrara, E. 2019. Perils and Challenges of Social Media and Election Manipulation Analysis: The 2018 US Midterms. In *Companion Proceedings of The 2019 World Wide Web Conference*, 237–247.

Del Vicario, M.; Bessi, A.; Zollo, F.; Petroni, F.; Scala, A.; Caldarelli, G.; Stanley, H. E.; and Quattrociocchi, W. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3): 554–559.

Del Vicario, M.; Scala, A.; Caldarelli, G.; Stanley, H. E.; and Quattrociocchi, W. 2017. Modeling confirmation bias and polarization. *Scientific Reports*, 7: 40391.

Elmas, T.; Overdorf, R.; Özkalay, A. F.; and Aberer, K. 2019. Lateral Astroturfing Attacks on Twitter Trending Topics. *arXiv:1910.07783*.

Ferrara, E. 2017. Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election. *First Monday*, 22(8).

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM*, 59(7): 96–104.

Flores-Saviaga, C.; Keegan, B.; and Savage, S. 2018. Mobilizing the trump train: Understanding collective action in a political trolling community. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*.

Fruchterman, T. M.; and Reingold, E. M. 1991. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11): 1129–1164.

Garrett, R. K. 2009. Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of Computer-Mediated Communication*, 14(2): 265–285.

Gleason, B. 2013. # Occupy Wall Street: Exploring informal learning about a social movement on Twitter. *American Behavioral Scientist*, 57(7): 966–982.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425): 374–378.

Hills, T. T. 2019. The Dark Side of Information Proliferation. *Perspectives on Psychological Science*, 14(3): 323–330.

Jamieson, K. H.; and Cappella, J. N. 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

Kitsak, M.; Gallos, L. K.; Havlin, S.; Liljeros, F.; Muchnik, L.; Stanley, H. E.; and Makse, H. A. 2010. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11): 888–893.

Lazer, D.; Baum, M.; Benkler, Y.; Berinsky, A.; Greenhill, K.; Menczer, F.; Metzger, M.; Nyhan, B.; Pennycook, G.; Rothschild, D.; Schudson, M.; Sloman, S.; Sunstein, C.; Thorson, E.; Watts, D.; and Zittrain, J. 2018. The science of fake news. *Science*, 359(6380): 1094–1096.

Lee, J. K.; Choi, J.; Kim, C.; and Kim, Y. 2014. Social media, network heterogeneity, and opinion polarization. *Journal of Communication*, 64(4): 702–722.

Liu, Y.; Liu, Y.; Zhang, M.; and Ma, S. 2016. Pay Me and I'll Follow You: Detection of Crowdturfing Following Activities in Microblog Environment. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 3789–3796.

Lotan, G.; Graeff, E.; Ananny, M.; Gaffney, D.; Pearce, I.; et al. 2011. The Arab Spring— the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, 5: 31.

Menczer, F.; and Hills, T. 2020. The attention economy. *Scientific American*, 323(6): 54–61.

Nizzoli, L.; Tardelli, S.; Avvenuti, M.; Cresci, S.; and Tesconi, M. 2021. Coordinated Behavior on Social Media in 2019 UK General Election. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media*, volume 15, 443–454.

Pacheco, D.; Hui, P.-M.; Torres-Lugo, C.; Truong, B. T.; Flammini, A.; and Menczer, F. 2021. Uncovering coordinated networks on social media: methods and case studies. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media*.

Pennycook, G.; and Rand, D. G. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7): 2521–2526.

Sasahara, K.; Chen, W.; Peng, H.; Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2020. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2018. The spread of low-credibility content by social bots. *Nature Communications*, 9(1): 1–9.

Sharma, K.; Zhang, Y.; Ferrara, E.; and Liu, Y. 2021. Identifying Coordinated Accounts on Social Media through Hidden Influence and Group Behaviours. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1441–1451.

Starbird, K.; and Palen, L. 2012. (How) will the revolution be retweeted? Information diffusion and the 2011 Egyptian uprising. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 7–16.

Stella, M.; Ferrara, E.; and 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.

Stewart, L. G.; Arif, A.; Nied, A. C.; Spiro, E. S.; and Starbird, K. 2017. Drawing the lines of contention: Networked frame contests within# BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, 1(96): 1–23.

Stringhini, G.; Wang, G.; Egele, M.; Kruegel, C.; Vigna, G.; Zheng, H.; and Zhao, B. Y. 2013. Follow the green: growth and dynamics in twitter follower markets. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, 163–176.

Tufekci, Z.; and Wilson, C. 2012. Social media and the decision to participate in political protest: Observations from Tahrir Square. *Journal of Communication*, 62(2): 363–379.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*, 359(6380): 1146–1151.

Wojcieszak, M. 2010. 'Don't talk to me': effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. *New Media & Society*, 12(4): 637–655.

Yang, K.-C.; Pierri, F.; Hui, P.-M.; Axelrod, D.; Torres-Lugo, C.; Bryden, J.; and Menczer, F. 2021. The COVID-19 Infodemic: Twitter versus Facebook. *Big Data & Society*, 8(1): 20539517211013861.

Yang, K.-C.; Torres-Lugo, C.; and Menczer, F. 2020. Prevalence of Low-Credibility Information on Twitter During the COVID-19 Outbreak. In *Proceedings ICWSM International Workshop on Cyber Social Threats*.

Yang, K.-C.; Varol, O.; Davis, C. A.; Ferrara, E.; Flammini, A.; and Menczer, F. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1): 48–61.

Yang, K.-C.; Varol, O.; Hui, P.-M.; and Menczer, F. 2020. Scalable and generalizable social bot detection through data selection. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, 1096–1103.

Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019. Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web. In *Companion Proceedings of The 2019 World Wide Web International*, 218–226.

Zarocostas, J. 2020. How to fight an infodemic. *The Lancet*, 395(10225): 676.