Title: Appendix for "COVID-19 misinformation and the 2020 U.S. presidential election"
Authors: Emily Chen (1), Herbert Chang (2), Ashwin Rao (1), Kristina Lerman (1), Geoffrey Cowan (2), Emilio Ferrara (1, 2)
Date: March 3rd, 2021
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix

A. Constructing the dataset

We found the most frequent hashtags, keywords, bigrams, and trigrams to understand the content of these topics and identified four broad narratives as discussed in Finding 1 and in Figures 1 and 2. Using keywords that best described these narratives, we then filtered both our COVID-19 and U.S. presidential elections dataset for tweets that contained at least one keyword from the primary-related keywords (Table 2) and the narratives of interest keywords (Table 3) and merged the two together. Because our COVID-19 dataset was specifically tracking COVID-19-related discourse, we felt it necessary to expand our subset of data to include the discussion on these narratives that were captured in our U.S presidential elections dataset to give us even more insight into how COVID-19 shaped primary discussion. This final dataset contained a total of 67,846,555 tweets, with 10,536,524 directly mentioning one of the COVID-19 related keywords, 5,900,737 referencing mail-in ballots, 1,283,450 tweets referencing mask-related discourse, and 619,914 tweets referencing lockdown measures.

B. Tagging public health misinformation

Upon tagging each user with one of the four classifications (Democrat and fact, Democrat and misinformation, Republican and fact, Republican and misinformation), we filter for tweets based on hashtags that are aligned with the different ideologies within the topic campaign and refer to them by their representative hashtags (#WearAMask, #MasksOff, #VoteByMail, #VoterFraud) throughout this paper.

The full list of hashtags aligned to each representative hashtag can be found in Table A1 below. We then identify the prevailing narratives present in each of the groups by examining the tweet n-grams.

Table A1. *Hashtags aligned with a specific representative hashtag.*

#WearAMask (n=45,108)	#MasksOff (n=7,236)	#VoteByMail (n=171,453)	#VoterFraud (n=67,488)
Wearamask	NoMasks	SaveThePostOffice	VoterFraud
Wearadamnmask	MasksOff	DontMessWithUSPS	NoMailinVoting
Maskup	MasksOffAmerica	MailinVoting	VoterIDNow
	NoMask	VoteByMail	MailinVoterFraud
		MailinBallots	VoterFraudIsReal
		SaveTheUSPS	DemVotebyMailScam
		SaveUSPS	
		USPSsabotage	
		VoteByMail2020	
		MailinBallot	
		USPSisEssential	

Note: For our mail-in ballot and masks-related subsets, we find the top 100 hashtags in each subset and isolate policy-stance related hashtags. We filter tweets based on these hashtags (case insensitive) to find a subset of tweets related to fact and misinformation views on masks and voting.

Table A2. *Number of tweets.*

	#WearAMask	#MasksOff	#VoteByMail	#VoterFraud
L/LL	8,626	364	52,343	1,592
C	624	117	3,529	630
R/LR	378	1,178	10,576	5,089

Note: We divide user views along their perspectives towards masks and mail-in voting and find the number of tweets for left (L), lean left (LL), center (C), lean right (LR) and right (R).

We find that for the subset of tweets that align with #WearAMask posted by Liberal users, the discourse encourages others to comply with regulations to wear masks. Some of the most frequent bigrams include "social distancing" and "wearing mask." We then look at tweets from Conservative users and find their conversation revolves around Donald Trump's decision to wear a mask and how this action can be used against the Democrats.

Figure A1. Screenshot of one of the tweets driving misinformation on mask discourse in both the Democratic and Republican parties.

However, when we look at the #MasksOff discourse, we find that regardless of party affiliation, both Conservatives and Liberals amplify misinformation messaging claiming that doctors believe that masks are adverse for one's health (an example of one such tweet can be seen in Figure A1).

For mail-in ballots, liberals tweeting #VoteByMail frequently mention "vote safely," "expand votebymail," and "wear mask," all of which suggest that Liberals are encouraging voting by mail as a means to remain safe during the COVID-19 pandemic. Conservatives are also voicing the same concerns, with mentions of "stay home," "social distancing," but also amplify their unhappiness regarding the Texas Supreme Court's decision to deny Democratic efforts to expand mail-in voting in Texas. On the other side of the spectrum, Liberals and Conservatives posting #VoterFraud-related tweets all reference a testimony given to the House Judicial subcommittee that supports the notion that a shift in mail-in ballots will increase voter fraud in the upcoming U.S. presidential election.

What we find there is that, for tweets supporting factual information, there is slight variance in topic coverage when we compare tweets from users in different parties that are engaging in the same information stance (misinformation versus factual information). However, when we examine misinformation content, there is homogeneity between what users from both parties are pushing on Twitter. This suggests that, for both mail-in ballot and mask-related discourse, both the left and right are susceptible to the same kinds of misinformation.

C. Limitations

It is difficult to compare survey reported political affiliations with political affiliation inferred through social media posts (Deb et al., 2019). Because our data set was filtered for keywords directly related to the 2020 US Democratic primaries, we see a significantly larger volume of tweets from Democratic tagged users, and a much smaller number of tweets attributed to Republican users. Thus, conclusions regarding Republican and Republican-leaning users' narratives were based on a small sample size of users.

We also note that Twitter's free streaming API only returns 1% of the total tweet stream. This means that we are not able to collect all of the tweets that are a part of the COVID-19 and Democratic primary-related discourse. However, the 1% sample still serves as a fairly accurate representation of the discourse. Twitter has also recently removed location data from a tweet's metadata, which means that we have had to infer user location based on the user reported location. These locations may not consistently be accurate, and we are unable to identify geolocation data for users who do not specify a location or users who fail to list a location from which we are able to extract location data.