Characterizing User Susceptibility to COVID-19 Misinformation on Twitter

Xian Teng[1], Yu-Ru Lin[1], Wen-Ting Chung[2], Ang Li[1], Adriana Kovashka[1]

[1]School of Computing and Information
[2]Department of Psychology in Education, School of Education
University of Pittsburgh
{xit22, yurulin, wtchung, ang125}@pitt.edu, kovashka@cs.pitt.edu

Abstract

Though significant efforts such as removing false claims and promoting reliable sources have been increased to combat COVID-19 "misinfodemic", it remains an unsolved societal challenge if lacking a proper understanding of susceptible online users, i.e., those who are likely to be attracted by, believe and spread misinformation. This study attempts to answer *who* constitutes the population vulnerable to the online misinformation in the pandemic, and what are the robust features and short-term behavior signals that distinguish susceptible users from others. Using a 6-month longitudinal user panel on Twitter collected from a geopolitically diverse network-stratified samples in the US, we distinguish different types of users, ranging from social bots to humans with various level of engagement with COVID-related misinformation. We then identify users' online features and situational predictors that correlate with their susceptibility to COVID-19 misinformation. This work brings unique contributions: First, contrary to the prior studies on bot influence, our analysis shows that social bots' contribution to misinformation sharing was surprisingly low, and human-like users' misinformation behaviors exhibit heterogeneity and temporal variability. While the sharing of misinformation was highly concentrated, the risk of occasionally sharing misinformation for average users remained alarmingly high. Second, our findings highlight the political sensitivity activeness and responsiveness to emotionally-charged content among susceptible users. Third, we demonstrate a feasible solution to efficiently predict users' transient susceptibility solely based on their short-term news consumption and exposure from their networks. Our work has an implication in designing effective intervention mechanism to mitigate the misinformation dissipation.

Introduction

Even before the WHO declared COVID-19 a pandemic with accompanying "massive infodemic", the issues of COVID-19 misinformation and how it ramifies with the pandemic has been one of the foci for the science community, industry, public health professionals, and the general public. Abundant media and research reports have shown how the surge of misleading and false information spreading online triggered a wide range of socially harmful behaviors (Ognyanova et al. 2020; Bridgman et al. 2020) that can jeop-

ardize the trust of authority sources, effectiveness of the epidemic control measures and further impact all sectors of a society. While a significant research effort has been spent on cutting the supply of misinformation – e.g., to fact-check or detect misinformation (Ball and Maxmen 2020), unless completely effective, cutting the supply alone will not solve the problem (Ball and Maxmen 2020). To intervene misinformation dissipation, there is a critical need in understanding users' (mis-)information behaviors.

Recent literature has shed light on the online misinformation phenomena. Several studies found that social bots are the main actors who spread low-credibility content, manipulated social media users' opinions and swayed democratic processes such as Brexit referendum and the US 2016 election (Shao et al. 2018; Ferrara et al. 2016; Bessi and Ferrara 2016). Apart from the bot influence, Grinberg et. al. (Grinberg et al. 2019) discovered that a handful of "supersharers" and "superconsumers" accounted for nearly 80% of fake news spread and consumption on Twitter during the past US presidential election. These studies together painted a picture of the key drivers in the fake news and misinformation campaigns that may influence political discourse and behaviors. In public health domain, the interlinking among misinformation and infectious diseases, such as the relationship with anti-vaccine attitude and other misconception, has been examined in other epidemics (Burki 2019). Since the COVID-19 outbreak, several studies have highlighted the ill effect of belief in false claims and its stemming on human's social and behavioral characteristics – for example, people with particular cognitive tendency and sociopolitical identities may be more vulnerable to misinformation (Lazer et al. 2020; Grinberg et al. 2019; Guess, Nyhan, and Reifler 2020). Despite the rich literature, there is still no clear picture for how misinformation penetrates the *ordinary* users' information processing in the life during a pandemic. Rather than focusing on a particular role such as bots or super-users, we seek to understand what distinguish users who are actively engage with misinformation from those less or rarely.

Who – bots or ordinary humans – constitutes the population vulnerable to the online misinformation in the pandemic? To inform the design of timely intervention, can we answer this without knowing their cognitive tendencies and sociopolitical identity? Relying on observable social media traces, we focus on three research questions:

- **RQ1.** To what extent do ordinary humans, compared to bots and super users, contribute to the COVID-19 misinformation spread? How do individuals engage with misinformation differently?
- **RQ2.** What are the robust features that distinguish susceptible users from those who have been resilient against the misinformation exposure?
- **RQ3.** Can users' susceptibility to COVID-19 misinformation captured by time-varying, situational factors? To what extent we can identify users' risk of sharing misinformation purely depend on short-term behavior traces?

To answer these RQs, we constructed a list of news sources flagged by journalists, fact-checkers and academia, and curated a longitudinal user panel on Twitter that were diverse in geography in United States. Our study period about COVID-19 misinformation covered the first 6 months of this pandemic. Our data also traced a user's timeline back to pre-pandemic period in order to understand the association between a user's past behaviors and misinformation engagement during this pandemic. Based on our data, we conducted a comprehensive analyses that compared user accounts by type (bots v.s. human-like users), by human-like users having different levels of misinformation engagement (RQ1) and behavioral characteristics (RQ2). Furthermore we exploited state-of-the-art deep learning models to perform real-time prediction purely based on short-term traces (RQ3). This work brings several unique contributions:

- **Analysis of misinformation engagement among the ordinary users.** Contrary to the prior studies on bot influence, our analysis shows that social bots' contribution to the misinformation sharing was surprisingly low, and majority of bots did not maliciously disseminate misinformation. On the other hand, *human-like* users' engagement with misinformation exhibit heterogeneity, where the sharing of misinformation was highly concentrated, but the risk of occasionally sharing and retweeting misinformation for average users remained alarmingly high.
- **Distinctive features correlated with users' susceptibility.** We identify features that distinguish users who are susceptible from those resilient to misinformation. Complementary to existing work, our results highlight the political sensitivity, activeness in tweeting and responsiveness to emotionally-charged content among susceptible users to COVID-19 misinformation.
- **Predictive models for identifying future risk from short-term behavior traces.** Given that our analysis suggesting the transient nature of users' dynamic misinformation behaviors, we propose a feasible solution, using an interpretable deep learning approach, to identify the future risk of misinformation sharing solely based on users' weekly behavioral and exposure traces.[1]

Related Work

COVID-19 Infodemic. To fight against the infodemic, researchers have been working to study the spread of misinformation on social media (Hameleers, van der Meer, and

---

[1]Codes at https://github.com/picsolab/COVID-misinfo

Brosius 2020; Papakyriakopoulos, Serrano, and Hegelich 2020; Memon and Carley 2020), the prevalence, predictors, causes, and consequences of belief in COVID-19 false statements or/and conspiracy theories (Lazer et al. 2020; Roozenbeek et al. 2020; Uscinski et al. 2020). It was found that citizens from US, UK, Netherlands and Germany experienced relatively high levels of mis-/disinformation – 4.88/4.46 on a 7-point scale – during the first stage of this pandemic (Hameleers, van der Meer, and Brosius 2020). Although platforms' moderation practices were effective in reducing false claims, certain issues (e.g, timeliness, magnitude and moderation bias) were identified (Papakyriakopoulos, Serrano, and Hegelich 2020). Studies have also shown that a substantial proportion of respondents indicated that they believe COVID-19 conspiracy theories, for example, Lazer et al. show that 7%-22% of US respondents believe 11 false claims about COVID-19 (e.g., coronavirus is created as a weapon in a Chinese lab, only older people get COVID-19), Roozenbeek et al. show that 22–23% respondents in UK, 26% in Ireland, 33% in Mexico and 37% in Spain agreed that coronavirus was engineered in a laboratory in Wuhan. Such susceptibility might lead to undesirable consequences as studies have suggested that belief in false claims is negatively related to vaccine acceptance and self-reported compliance with guidelines (Roozenbeek et al. 2020; Lazer et al. 2020; Bridgman et al. 2020). Our work is different from prior work in several aspects: first, our study leverages large-scale social media traces instead of survey data to examine online users' susceptibility, second, our study is not limited to the several most popular COVID-19 conspiracy theories, instead we study the broad spectrum of online COVID-19 misinformation; third, our work considered a broad spectrum of online actors including bots, superusers and average users to understand to what extent and how (strategies) they contributed to misinformation spread.

The Fake-News Phenomena on Social Media. Fake news on social media has become a public concern. Existing work that are most related to ours include bots/trolls influence in promoting unreliable contents (Ferrara et al. 2016; Ferrara 2020; Badawy, Lerman, and Ferrara 2019), and predictors for users' susceptibility, e.g., such as demographics, socio-psychological attributes, political learning, news diets (Grinberg et al. 2019; Allcott and Gentzkow 2017). It has been found that bots occupy a relatively high proportion (nearly 15%) of Twitter accounts in 2016 US presidential election (Varol et al. 2017; Bessi and Ferrara 2016), they are more active in posting and likely to mention influential accounts to promote low-quality content, therefore succeed to spread information into human population (Shao et al. 2018). In addition to bots, studies have suggested that certain types of human users are more vulnerable to misinformation, i.e., people who are older (Grinberg et al. 2019), less educated (Allcott and Gentzkow 2017), maintaining a more conservative news diet or/and supporting Trump (Grinberg et al. 2019; Guess, Nyhan, and Reifler 2020). Contrary to prior work concerning bot impact, this study finds that humans (not bots) have been heavily engaged with COVID-19 misinformation on Twitter. This finding echos

a recent study (Silva et al. 2020), indicating that 8.5% were bots among those who tweeted COVID-19 misinformation, however, this study discarded all duplicated retweets which might fail to account for bot activities in resharing COVID-19 misinformation. Furthermore, we observed that humans exhibited heterogeneity in their misinformation behaviors in terms of volume, strategies and temporal dedication. Regarding susceptibility predictors, our results highlight the responsiveness of susceptible users to emotionally- and politically charged content.

Combating Mis-/disinformation. The computing community is engaging in the design of advanced methods to automatically combat mis-/disinformation on social media, including bot detection (Ferrara et al. 2016), fake news detection (Shu et al. 2017) and cascade intervention (**?**). Bot detection systems can be divided into three classes: techniques based on social network, systems leveraging human intelligence, and machine learning models based on discriminative features (Ferrara et al. 2016). However, the deployment of bot detection systems is undermined by the demand of significant amount of user-level information, and the challenge to account for bots' continuously changing behaviors (Ferrara et al. 2016). Fake news detection techniques typically leverage information from news contents (texts, images and videos) and social contexts (user profiles, post responses, social networks) (Zhou et al. 2019). A major issue is that fake news are constantly being produced in large scale following emerging and time-critical events, therefore easily spread to large audience without early warning or/and being fact-checked. Cascade intervention leverages the understanding of diffusion dynamics to monitor a small set of influential nodes or intercept certain propagation paths to limit the spread of misinformation, yet it requires an overview of social network graphs. Contrary to prior efforts, we propose a different task to battle misinformation, i.e., to proactively foresee a user's near-future susceptibility by solely using its short-term behavior traces. Our aim is to develop a situational and social-aware tool to catch up risk early before misinformation are further shared to larger population. For this purpose, we develop a model exploiting state-of-the-art deep learning techniques, which is able to capture the correlations of domains as well as dynamics of news ecosystem leveraging users' temporal cosharing/coexposure behaviors.

Data & Study Design

This section describes our data collection and processing, as well as crucial definitions and measurements.

Data

Flagged News Sources. We adopted a list of flagged news sources as proxy of misinformation following suggestions by prior work (Lazer et al. 2018), since using fact-checked articles might end up biasing towards a small set of "popular" or/and "fact-checkable" news stories. Our flagged news source dictionary was constructed by referring to two datasets. (i) We first included the news sources curated by MediaBias/FactCheck (MBFC hereafter) under two categories – conspiracy-pseudoscience and questionable sources

(accessed on April 19, 2021), where the the conspiracy-pseudoscience category "may publish unverifiable information that is not always supported by evidence", and the questionable source "exhibits one or more of the following – extreme bias, consistent promotion of propaganda/conspiracies, poor or no sourcing to credible information, a complete lack of transparency and/or is fake news." We used the most updated version of MBFC news sources in April 2021 (prior to our paper submission), to included the most up-to-date information about websites that published COVID-19 misinformation during our study period.[2] (ii) We also included several pre-existing lists of websites (Grinberg et al. 2019), including three sets of publishers released by fact-checkers and journalists (i.e., published by Buzzfeed News, Politifact, and FactCheck.org), and two sets of problematic domains labelled by scholars (Guess, Nyhan, and Reifler 2018; Grinberg et al. 2019). Following Grinberg et al., we only included the domains if labeled as black, red or orange as these colors reflected annotators' stronger affirmation regarding a flawed editorial process (e.g., little regards for the truth, negligent or deceptive). In total, there were 1528 flagged news sources in our dictionary, and 527 of them appeared in our tweet data.

User Panel Construction. To identify users contributed to COVID-19 information sharing, we first referred to a public data of COVID-19 Twitter IDs between January 28 and April 24 2020 (Chen, Lerman, and Ferrara 2020). To obtain a user panel that were diverse in geography and ideological context, we selected four cities (with a population larger than 250K people) spanning from liberal to conservative based on conservatism scores estimated by (Tausanovitch and Warshaw 2014), including San Francisco CA, New York City NY, Houston TX and Nashville TN.[3] To understand users' social network properties, we randomly selected 100 seed users in each city, and snowball their friends to construct a social network. Specifically, the seeds were guaranteed to be active human accounts – i.e., they were not detectable bots or organizational accounts (detected using tools discussed below) and they had posted at least two URLs prior to April 24, 2020. We ensured that half of the seeds shared a proportion larger than 60% of URLs from flagged sources while the other half shared less than 40% URLs from those sources, so as to capture both resilient and susceptible users (formal definitions can be found in section). Starting from the seeds, we obtained up to 5000 users IDs they have been following (known as their "friends") as of June 2020. This step was run for two steps (therefore two-step snowball sampling) and resulted in more than 9.5 millions unique user IDs. We kept the accounts having at least one friend and one follower in the social network to make our panel in a reasonable size and facilitate our subsequent social network analyses. Then

---

[2]We manually checked MBFC's analysis reports for flagged websites and found at least 143 websites were labeled by MBFC to be the sources spreading COVID-19 misinformation.

[3]These four cities were chosen to achieve a trade-off between diversity and availability of user samples. Our user distribution across cities looks similar to the distribution obtained from the above public COVID-19 data crawled using keywords (Chen, Lerman, and Ferrara 2020).

we crawled the most recent 3000 tweets posted by each user as of June 2020. In order to exclude accounts that were created after outbreak to be used to run COVID-19 related campaigns, we only kept users who registered their accounts before November 1 2019. We also removed inactive accounts who have not posted any tweets between December 1 2019 and June 1 2020. To guarantee we cover a user's complete activities after outbreak, we only kept users whose oldest tweet in our data was published prior to December 1 2019. Eventually, we ended up with 531,865 users and approximately 1.4 billions tweets.

Location Extraction.   We extracted users' locations based on the location field within a user object. We matched the text against a list of cities, counties, states and city/state abbreviations in the US. For example, we extracted location "Winslow, Arizona" from the text "on a corner in Winslow, Arizona," and "NYC, NY" from the text "NYC, New York." In total, 287,784 users didn't disclose any city or state information, 244,081 users provided state information and 168,788 users provided city information.

COVID-19 Tweets.   As our study is in the COVID-19 context, we particularly identified COVID-19 tweets by searching for any of the COVID-19 related keywords constructed in (Chen, Lerman, and Ferrara 2020) within text, URLs, hashtags, mentions and screen names.

URL Expansion.   Many users leveraged shortened links to share long URLs to maintain the maximum number of characters for posts. For the URLs included in COVID-19 tweets, we followed all redirects and obtained the final website domain. But for the URLs in COVID-19 irrelevant tweets, it would be impractical to expand them all within a reasonable time period. Therefore, we recovered a subset of shorted URLs from known URL shorteners to the original domains (e.g., "youtu.be" to "youtube.com", "wapo.st" to "washingtonpost.com"), and merged with long URLs to be used in our subsequent analysis.

Definitions and Measurements

Defining Study Period.   As described in data section , our data traces a user's data before outbreak as well as during the first 6 months of this pandemic until early June 2020. Since the first human cases of COVID-19 were reported to be identified in Wuhan, China in December 2019, which triggered large-scale online discussion on social media ever since, we defined December 1 2019 to June 1 2020 as our COVID-19 misinformation study period $\mathbf{P}_{cov}$. We considered the time prior to December 1 2019 as pre-pandemic period $\mathbf{P}_{pre}$.

Defining Susceptible/Resilient Users.   We rely on the sharing activities of users to study misinformation engagement, defining a *share* as publishing a post (e.g., tweet, retweet, quote or reply) that included a URL redirecting to a page outside of Twitter. A single post consisting of five external URLs would be counted as five shares. If the post were an original tweet, we would have five *original shares*. If one contained URL were from one of the flagged sources, we would call it a *problematic share*; We generally called URLs from unflagged sources as *other shares*. Accordingly,

a *susceptible user* is defined as an account who had at least one problematic share, and a *resilient user* as who didn't share any URLs from flagged sources. Furthermore, we introduced a susceptibility adherence metric as level of engagement with content from flagged sources, measured by the number of problematic shares. If restricting our scope to COVID-19 tweets, we have COVID-19 sharers, COVID-19 problematic shares, COVID-19 susceptible and resilient users. We will omit the term "COVID-19" in subsequent discussion if no special instruction. In summary, there are 265,859 sharers – 88,160 are susceptible and 177,699 are resilient, who made 5,515,408 shares – 560,788 are problematic shares and 4,954,620 other shares.

Identifying User Types.   We utilized a machine learning tool called *Humanizr* (McCorriston, Jurgens, and Ruths 2015) to automatically identify organizational accounts (e.g., institutions, corporations etc). In total, 764 organizational accounts were located and excluded in our analysis. In order to further distinguish bot-like and human-like accounts, we used a bot detection tool system *Botometer* to assign a continuous score to each user (Varol et al. 2017). To avoid time misalignment, our Botometer analysis was conducted immediately after we had collected Twitter user IDs between June 14 and July 9.

Estimating News Medias Bias.   We inferred a news media's political alignment by aggregating four distinct but overlapping datasets: two lists of news sources containing 500 and 224 sites from prior research (Bakshy, Messing, and Adamic 2015; Grinberg et al. 2019), a list of 1658 news medias labeled by Allsides.com, and a list of 312 news sites curated by MediaBias/FactCheck. Specifically, Grinberg et al. and Bakshy, Messing, and Adamic assessed a media's political alignment as the proportion of registered Republicans and Democrats who engaged (exposed or shared) with the source, thus a continuous score (-1 is most left, +1 right) is assigned to each website. Allside's media bias were obtained through a hybrid approach including editorial review, blind survey, third-party analysis and independent research. MediaBias/FactCheck's ratings were average scores by considering aspects including biased wording/headlines, factual/-sourcing, story choices and political affiliation. The later two use categorical labels (left, left-center, center, right-center and right). In order to address the inconsistency of four lists, we employed an iterative imputation method "missForest" to impute missing values in four datasets (Stekhoven and Bühlmann 2012). In total, there were 2163 news medias in our analysis.

Identifying Right/Left-leaning Hashtags.   Twitter users include hashtags in their user profile descriptions to connect with ideologically similar users and express support for movements/politicians. We extracted the most popular hashtags that were present in at least 100 users' descriptions. We identify a hashtag as ideologically relevant if it either supports or criticizes political leaders, their statements and campaigns (#trump, #maga, #bluewave). To label a hashtag, we used Twitter's search engine to retrieve relevant tweets, people, photos and videos, then examined the top ten results in
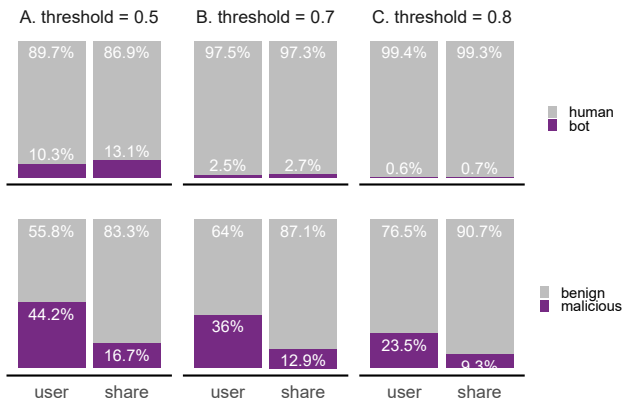
Figure 1: Prevalence of social bots estimated by varying bot score thresholds (A-C). Upper bars indicate separation of humans v.s. bots (left) and shares attributable to them (right). Lower bars indicate separation of benign v.s. malicious bots and shares attributable to them.

each of those categories. In total, there were 83 right-leaning and 69 left-leaning hashtags used in our analysis.

Analysis & Results

This section reports our analysis and experimental results for three research questions.

RQ1. To What Extent and How Differently Users Engage with Misinformation?

Bots' Contribution. First, we computed the proportion of social bots among all COVID-19 sharers. Following prior studies (Varol et al. 2017; Grinberg et al. 2019), we labelled bots at thresholds 0.5 and 0.7: an account was considered to be a bot if its Botometer score were above 0.5 or 0.7. We added another threshold at 0.8 to achieve a posterior probability of 50% (implying that with half chance an account having a score $\geq$ 0.8 is a bot) for sensitivity test. Fig. 1 reveals that at threshold 0.5 approximately 10.3% of misinformation sharers are bots, who are responsible for 13.1% of problematic COVID-19 shares. The proportion of bots is even smaller – 2.5% and 0.6% – if setting threshold to 0.7 and 0.8, accordingly the percentage of associated shares drops to only 2.7% and 0.7%. Furthermore, we calculated the decomposition of different types of bots – malicious v.s. benign. A bot was considered to be *malicious* if it had at least one problematic share, otherwise *benign*. We find that a large proportion of bots are actually benign – the fraction is 55.8% at threshold 0.5, 64% at threshold 0.7 and nearly 76.5% at threshold 0.8. By examining the decomposition of shares made by the two bot types, we find that benign bots are more active than malicious ones in sharing: at threshold 0.5, malicious bots only account for 16.7% of all bot shares while benign bots are responsible for up to 83.8% of all bot shares; The activity difference is even larger when increasing the threshold to 0.8, i.e., 9.3% of shares (note that not all of them are problematic shares) are from malicious
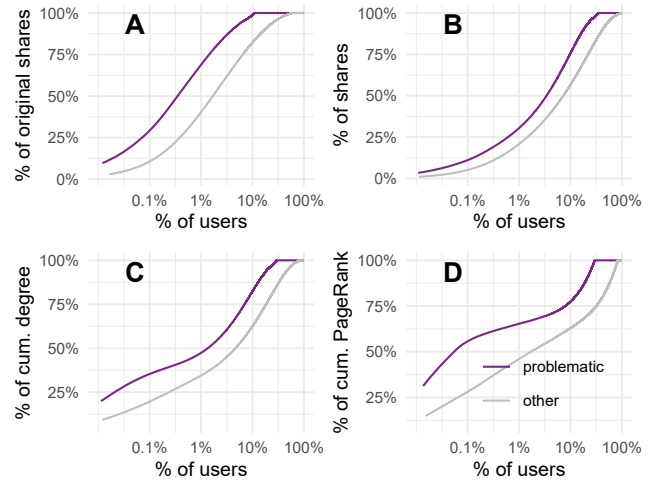

Figure 2: Concentration of misinformation engagement. Empirical cumulative distribution functions (ECDF) for number of shares in original tweets (A), number of shares in both original posts and retweets (B), degree and PageRank in retweet network (C,D). The x-axis represents the percentage of users responsible for a given percentage (y-axis) of cumulative values. Purple (gray) lines represent problematic (other) shares.

bots while nearly 90.7% shares from benign bots (therefore not problematic). To summarize, we find that in the COVID-19 context the prevalence of social bots among problematic sharers is relatively low, and surprisingly benign bots account for a very large fractions of presence and sharing activities. These findings in turn suggest that not bot but humans were mainly engaging with misinformation during our study period.

Superusers' Contribution. In subsequent analysis we excluded bot-like accounts using the lowest threshold 0.5 to guarantee that the remaining human sample do not contain bots. To test our hypothesis that the engagement with COVID-19 misinformation is heavily concentrated on a small fraction of core sharers, we computed four distinct metrics – number of original shares, number of shares, the degree centrality and PageRank in retweet network. The retweet network was constructed as follows: if a user A retweeted B's post that was COVID-19 related and contained external URLs, we connected a directed edge from A to B. The former two metrics capture a user's level of sharing activity, while the latter two characterize a user's influence in information spread. It is revealed that merely 1% of susceptible users account for almost 75% of original shares (Fig. 2A), and 10% of susceptible users account for 75% of shares (Fig. 2B). Fig. 2C,D show that super influencers are present in COVID-19 retweet network – if we consider degree and PageRank as indicators of "wealth", a 1% of top-ranked susceptible users hold 50% and 60% of the cumulative wealth in the sample. The concentration phenomenon is more pronounced in problematic sharing (purple) compared to the sharing of other contents (gray).
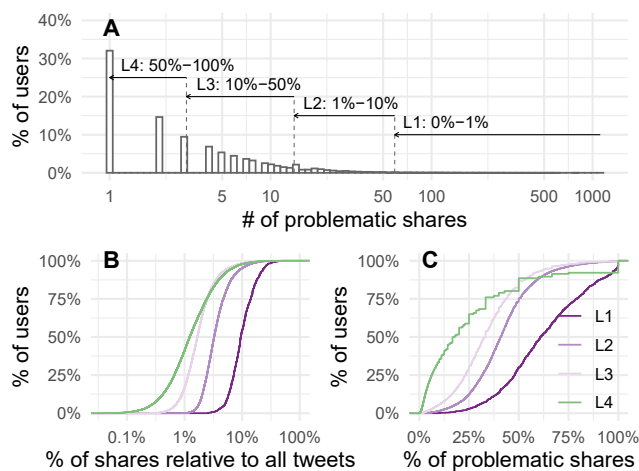
Figure 3: Dividing susceptible users into four levels. (A) Distribution of number of problematic shares and four levels L1-L4. (B-C) ECDF for the percentage of shares relative to COVID-19 tweets and the percentage of problematic shares relative to all COVID-19 shares. The y-axis indicates the percentage of users who have less than or equal to a certain value (x-axis).
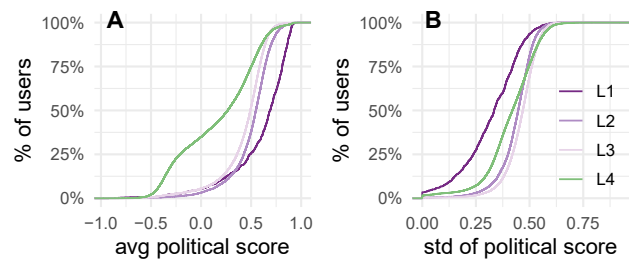

Figure 4: Consumption of news media for L1-L4 users. (A) Average political score of consumed news media. (B) Standard deviation of political scores.
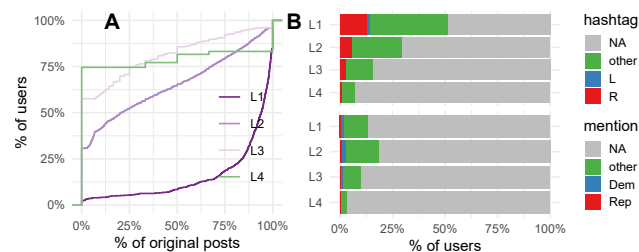

Figure 5: Tactics used in sharing COVID-19 misinformation for L1-L4 users. (A) ECDF for precent of original shares. (B) Percent of users who included hashtags and mentions in problematic shares.

Heterogeneity of Humans' Misinformation Behaviors. The preceding analysis suggests that users engaged with COVID-19 misinformation at different levels. We furthermore answered the question: how different are users who are strongly adherent to flagged sources, i.e., high-adherent users, compared to those who are weekly adherent? For this purpose, we divided susceptible human-like users into four levels: L1 included top 1% users who had the largest number of problematic shares, L2 between 1% and 10%, L3 between 10% and 50%, and L4 between 50% and 100%. We found that nearly half of susceptible users only shared no more than two problematic URLs, raising the alarm that the risk of occasionally being exposed to or/and sharing misinformation is relatively high (Fig. 3A). Fig. 3B,C reveal that strong-adherent users (in contrast to weak-adherent users) have a stronger preference in using URLs when posting COVID-19 tweets, as the percentage of shares normalized by the number of COVID-19 tweets is relatively high (Fig. 3C); They have a stronger interest in sharing information from flagged sources as the percentage of problematic shares normalized by number of shares is relatively high (Fig. 3D). Specifically, we studied their differences in four aspects as follows.

(1) News source partisanship. We computed the average and standard deviation of political scores of news sources shared by a user. A larger (smaller) score indicates that the user likes to share news from right-leaning (left-leaning) media sources, a larger (smaller) standard deviation implies that the user has broader (narrower) range of news preferences regarding politics. Fig. 4A,B show that high-adherent users are more likely to consume right-leaning news sources and hold narrower political news diet compared to individuals from L3-L4 groups.

(2) Sharing tactics. We measured three possible tactics in

sharing misinformation: (i) proactively introducing external misinformation into Twitter; (ii) exploiting hashtags – particularly left/right-leaning hashtags – to increase visibility of the shared misinformation; (iii) mentioning other accounts – particularly Democratic or Republican leaders (senators and house representatives) – while sharing misinformation. For (i), Fig. 5A reports the fraction of original shares relative to all problematic shares, suggesting that L1 users were more likely to introduce misinformation into Twitter, whereas more than half of L3-L4 individuals simply reshared the misinformation that had already existed. For (ii-iii), we calculated the percentage of individuals within each of the groups who had included certain hashtags or mentioned others in any of their problematic shares (Fig. 5B). We found that L1 users were prone to include hashtags – especially right-leaning hashtags – to increase visibility, whereas this association is not observed in the usage of left-leaning hashtags; Users rarely mentioned others while sharing misinformation, there is no significant association observed between mentions and misinformation engagement.

(3) Social network characteristics. We investigated to what extent high/weak-adherent users were located at important positions in retweet networks. Fig. 6A shows that L1 users have larger values of degree in retweet network (similar pattern is also observed for PageRank, we omit its plot to save space), suggesting that high-adherent users played a more successful role in spreading information into social network. One might suspect that perhaps they have more au-

dience or/and were retweeted more often. So we compared the number of followers (Fig. 6B) and calculated the average number of times being retweeted per problematic share for users in L1-L4 groups (Fig. 6C). We found that high-adherent users didn't hold much structural advantages based on followers count, but they were more successful in misinformation diffusion process as L1 users were retweeted most often among all groups, approximately 80% of L3-L4 users were not even retweeted by anyone in our data.

(4) Exposure of misinformation from social friends. Before a user shares a URL at time t, they might have seen the same URL from their friends' tweets shared at an earlier time t' ($t' < t$). We considered a potential exposure for an ego user to be any tweet shared by one of his/her friends, that includes the same URL he/she has shared. If multiple matches were found, we consider the most recent one $\Delta t = t - t' < \tau$ and break down the range of τ into different values ($\tau = 1, 6, 12, 24, 48, 72$, and > 72 hours). If a share is successfully matched to a preceding exposure (PE), we call it a PE share, otherwise a non-PE share. We then consider the fraction of PE shares, i.e., the number of shares of type c that has a PE / the number of all shares of type c, for $c \in \{$problematic, non-problematic$\}$.

Although a matched preceding exposure does not sufficiently imply a causal influence of sharing, this analysis allows us to capture the patterns of temporal ordering of exposure and sharing behaviors. Fig. 7A reports the percentage of PE shares for both types of shares (problematic and non-problematic) at different τ's. It shows that problematic shares from ego users, compared to non-problematic shares, are more likely to have a PE from friends, suggesting that misinformation might prone to spread through follow-followee relationship than reliable information (Vosoughi, Roy, and Aral 2018). Fig. 7B,C shows the percentage of PE shares ($\tau \geq 0$) for problematic and non-problematic respectively, breaking down by L1-L4 users. Both figures suggest that average users (L2-L4) were more likely to reshare information from PE from their friends than L1 users.

(5) Temporal variation of misinformation engagement. Individuals' misinformation behaviors exhibited temporal variation, i..e, a user might share problematic URLs in certain weeks while not in others. We consider a user's status during a week as *SH* if he/she shared misinformation, and *NS* if didn't share any misinformation. Then we calculated the proportion of users in each of L1-L4 groups flip status across consecutive two weeks over a 3-month period between March and May. We find that on average L1 users have the highest tendency to maintain SH status (70%±15%), while L4 users have the highest tendency to maintain NS status (80%±4%). In contrast, L2/L3 users were more likely than L1/L4 users to change status, for example, 14%-17% of them flipped their status while this number is only 5% for L1 and 10% for L4.

Overall, we find that users who heavily engaged with misinformation were more likely to share conservative websites, more active in introducing misinformation into the platform rather than simply forwarding it, and more strategic in promoting its spread. Average users, in contrast to those who heavily engaged with misinformation, have a tendency to
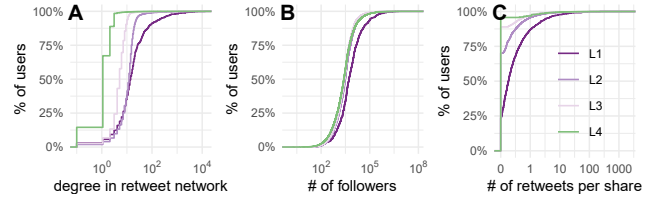

Figure 6: Social network properties for L1-L4 users. (A) ECDF of a user's degree in retweet network. (B) ECDF of an individual's followers counts. (C) ECDF for the number of retweets per problematic share.
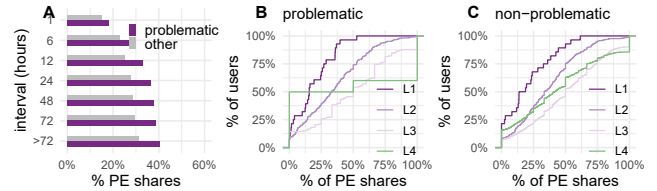

Figure 7: Exposure analysis from social context. (A) Fraction of shares that had a preceding exposure for both problematic and other-type information. (B) Fraction of problematic shares that had a preceding exposure for L1-L4 users. (C) Fraction of non-problematic shares that had a preceding exposure for L1-L4 users.

reshare information after their friends' sharing. The risk of occasionally sharing misinformation for average users remained high.

RQ2. Discriminative Features

RQ1 overviews to what extent as well as how various actors engaged in spreading misinformation, which reveals the heterogeneity among the susceptible population. In RQ2, we particularly zoom into the major subset of ordinary susceptible users, investigating the contexts and behaviors – in contrast against resilient users – to better understand the predictors that correlate with being susceptible to misinformation.

Comparison Groups. To obtain a panel of ordinary individuals we first excluded organizational accounts, bots (at threshold 0.5), and superusers (L1 group), we kept accounts with city information to further guarantee they are genuine users. We then further cleaned our samples by removing IQR outliers (in terms of friends count, followers count and number of tweets posted per day). We limited our samples to those who have posted at least 10 tweets in $\mathbf{P}_{pre}$, 10 tweets in $\mathbf{P}_{cov}$ and 4 COVID-19 shares. We examined each panelist's misinformation behaviors in $\mathbf{P}_{pre}$ and $\mathbf{P}_{cov}$ to assign two types of labels accordingly. There are 63,371 users in our final penal – 15,830 of them are *COVID-19 susceptible* among which 77% were *pre-pandemic susceptible*, and 47,540 *COVID-19 resilient* among which 65% were *pre-pandemic resilient*. For brevity, we denote COVID-19 susceptible/resilient users as CSUs/CRUs, pre-pandemic susceptible/resilient users as PSUs/PRUs below.

Features	ALL	PSU	PRU
Prev. suscept.	1.1***		
Swing county	0.0	0.1	-0.0
Republican county	0.2***	0.2***	0.2**
% leader interact (Rep)	10.0***	9.5***	9.1***
% leader interact (Dem)	4.0***	4.7***	3.5***
% retweet (Rep)	0.4**	0.3**	0.3
% retweet (Dem)	-0.8***	-0.8***	-0.7***
% reply (Rep)	-0.5**	-0.2	-1.2***
% reply (Dem)	0.6***	0.4*	0.6**
% mention (Rep)	-0.4***	-0.4**	-0.2
% mention (Dem)	-0.3**	-0.2	-0.4**
Sent. reply (Dem)	-0.2**	-0.1	-0.2*
Sent. reply (Rep)	0.1*	0.1	0.2
Sent. mention (Dem)	-0.2***	-0.2**	-0.2**
Sent. mention (Rep)	0.1	0.1	0.0
Profile hashtags (R)	1.0***	0.7***	2.8***
Profile hashtags (L)	0.4***	0.3***	0.5***
# tweets / day	5.9***	5.7***	6.3***
% interval < 1 hour	1.7***	1.3***	2.7***
% interval > 1 day	-1.7***	-2.4***	-0.6*
% tweets incl. hashtags	-1.2***	-1.3***	-1.0***
% tweets incl. mentions	0.4***	0.2	0.7***
% tweets incl. urls	0.2**	0.4***	-0.1
log(#follower/#followee)	-0.1**	-0.0	-0.2***
Age (months)	-0.2***	-0.3***	-0.0
Account verified	-0.5***	-0.5***	-0.3***
Positive emotion	-0.8***	-1.2***	-0.1
Negative emotion	3.2***	2.9***	3.3***
Avg. political score	2.1***	2.6***	0.5***
N	44925	25067	19858
AIC	36866	23486	12848
Null deviance	55880	34584	15313
Residual deviance	36802	23424	12786

*p<0.1, **p<0.05, ***p<0.001

Table 1: Coefficients from classifying COVID susceptible v.s. resilient users among (i) all users, (ii) PSU: the subgroup who used to be susceptible prior to this pandemic, and (iii) PRU: the subgroup used to be resilient.

Features. We examined a series of contextual variables. **(D1) Pre-pandemic susceptibility** describes whether a CSU/CRU used to be susceptible or not. **(D2) Geopolitical environment** characterizes the political climate of the county a user lives in using returns of US 2000-2016 presidential elections, e.g., a swing county or not, a Democratic or Republican county. **(D3) Political sensitivity** contains a user's tendency to interact with political leaders of Republican or Democratic party, the sentiment expressed in the interactions with them, as well as the variables indicating whether a user used left/right-leaning hashtags in profile descriptions. For leadership interaction, we examined the overall proportion of interactions with leaders from one party among all interactions with others, as well as the percentage of different types – retweet, mention and reply – among leadership interactions; for sentiment, we calculated the median score (between -1 and +1) along with each type of lead-

ership interactions. **(D4) Activeness** characterizes a user's activity pattern from distinct aspects such as tweeting intensity (tweeting rate and intervals), platform engagement (the proportion of using hashtags, URLs and mentions in tweets, ratio of followers to followees) and credibility (account age, verified or not). **(D5) Emotions** are based on LIWC lexicons by computing the user-level proportion of tweets containing tokens from each lexicon category. **(D6) News diets** captures the average political alignment score for news sources consumed by a user. **(D7) Social exposure** include the fractions of right (political score ≤ -0.2), left (political score ≥ 0.2) and neutral friends in social network.

Findings and Insights. Table 1 shows logistic regression analyses from three models: column "ALL" corresponds to the model that includes pre-pandemic susceptibility as one of the predictors, while column "PSU" and "PRU" correspond to the one that classifies CSUs/CRUs among the subgroup of users who used to be susceptible (PSU) and resilient (PRU) prior to this pandemic, respectively. It reveals that the misinformation engagement was positively associated with being previously susceptible (D1), coming from Republican counties (D2) and having conservative news diets (D6). Interestingly, in overall CSUs are more likely than CRUs to interact with political leaders from both sides (p<1e-3) – Democratic and Republican, as well as to insert partisan hashtags (p<1e-3) – both left and right – in their profile descriptions (D3). When we further examined interaction types, we found that CSUs have a stronger tendency to reply to rather than to retweet the posts of Democratic leaders. In terms of sentiment, we observed that CSUs like to express greater negative sentiment towards Democratic leaders, particularly when they mention or reply to a Democratic leader. Regarding activeness (D4), CSUs are more likely than CRUs to tweet frequently with shorter intervals, have accounts being unverified, but less likely to include general hashtags in tweets. CSUs have a tendency than CRUs to tweet with words showing negative emotions (D5). In terms of social exposures (D7), we found that CSUs have significantly more right-leaning friends (K-S D=0.56, p<1e-3), fewer left-leaning (K-S D=0.54, p<1e-3) and neutral friends (K-S D=0.19, p<1e-3) compared to CRUs.

The positive association of sharing misinformation with conservative news diets, geopolitical environment and social exposures can be linked to prior literature, as existing work have shown that conservatives are more likely than liberals to believe conspiracy theories (Enders and Smallpage 2019), tolerate misinformation from politicians (Roets et al. 2019), distrust mainstream news media outlets and heavily consume hyper-partisan right-wing media outlets which lack fact-checking or editorial norms (Marwick and Lewis 2017). Besides, our findings regarding COVID-19 susceptible users' overall political sensitivity and negative emotions are in alignment with prior evidence that individuals who likely to engage with flagged sources are also highly engaged with political news (Grinberg et al. 2019), and conservatives are more attracted than liberals by negativity and respond to various controversial issues with negative emotions (Inbar, Pizarro, and Bloom 2009). Our study high-

lights the patterns of susceptible users interacting with political figures from the opposing party: susceptible users are more likely to reply Democratic leaders instead of retweeting them (avoid endorsement or propagation). Besides, they choose to write negative words by replying to and directly mentioning Democratic leaders to express their opinions. These findings align with prior studies about adversarial interactions against candidates for the U.S. House of Representatives (Hua, Ristenpart, and Naaman 2020; Hua, Naaman, and Ristenpart 2020), which capture the cases of Twitter users' hostility using misinformation to attack and to undermine the legitimacy of candidates from opposing party, and those highly adversarial users express partisanship bias in their profiles. Furthermore, our results regarding activeness features can be explained by the difference in digital strategies of left/right-wing actors. Studies have shown that the left tend to directly rely on platforms to distribute messages, such as "hashtag activism" (Freelon, Marwick, and Kreiss 2020), by contrast the right believe that the "Big Tech" platforms are biased against them and they manipulate those platforms to amply their messaging to larger audience, e.g., gaming Twitter's trending topics feature, using fake accounts, leveraging partisan hashtags to find ideologically similar users (Marwick and Lewis 2017). Therefore, CSUs exhibited higher tweeting rate, shorter intervals, accounts being unverified, less general hashtags in tweets but more partisan hashtags in profiles, compared to CRUs.

RQ3. Situational Prediction of Susceptibility

Our results in RQ1 suggest that a user's susceptible state keep changing over time (section 4.1.3 (5)), as well as possibly triggered by prior exposure of (mis)information from the his/her the continuously evolving social environments (section 4.1.3 (4)). Therefore, in RQ3 we ask the question: can we develop an situational, social-aware model to predict a user's near-future susceptibility based on his/her recent sharing activities and exposure history from friendships? This task is particularly crucial in practical settings where we need to distribute attention resources to at-risk users and catch up misinformation early before they are further shared to a larger population.

Prediction Task and Model. In our prediction task, the target variable takes value 1 if a user, at a certain time point, shares misinformation (i.e., any URLs from flagged news sources) in the subsequent 7 days, 0 otherwise. The inputs include a sequence of website domains shared by the user during the past 2 weeks, called *share sequence*. Furthermore, to capture the possible trigger from social context exposure, we also generate another sequence of URLs shared by its 1-hop friends, called *exposure sequence*. We developed an interpretable deep learning model exploiting Transformer encoder to process sequential inputs (Vaswani et al. 2017), where the intrinsic interpretability comes from the design of a linear format, i.e., each domain's contribution to the final prediction could be captured by a corresponding coefficient (illustrated in Fig. 8). For comparison we implemented a logistic regression (LR) classifier that takes the counts of news sources encountered in the past as feature

	accuracy	precision	recall	F1
LR	0.69 (3e-3)	0.66 (4e-3)	0.76 (2e-3)	0.71 (3e-3)
CNN	0.71 (3e-3)	0.70 (1e-2)	0.75 (4e-3)	0.72 (1e-2)
GRU	0.71 (5e-3)	**0.70 (1e-2)**	0.75 (3e-2)	0.72 (1e-2)
Ours	**0.72 (4e-3)**	0.69 (7e-3)	**0.79 (1e-2)**	**0.74 (4e-3)**

Table 2: Model performance with average and standard deviation obtained from five trials of experiments.
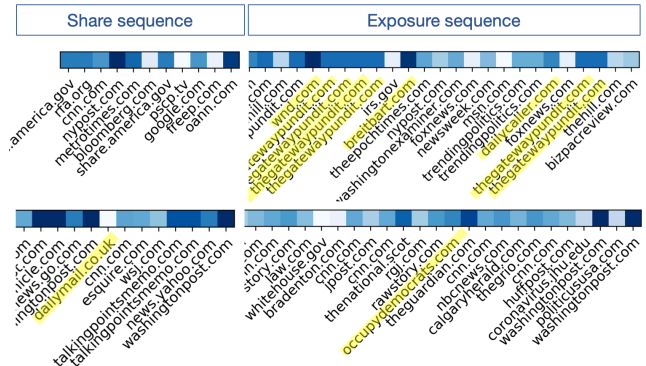

Figure 8: Contribution heatmap of domains from past share/exposure sequences. Darker color indicates larger contribution, flagged sources are highlighted.

vectors, and two neural network baselines based on CNN and GRU. A detailed description about our model is in Appendix (see Fig. 10), experimental settings in Appendix .

Performance and Interpretability. Table 2 reports prediction performance: the overall result suggests that we are able to make reasonable real-time predictions simply based on users' short-term traces; besides, our model and other neural network baselines outperform the simple LR classifier. Fig. 8 shows two case study to illustrate the interpretability of our model, as it calculates a contribution weight to each of the past shared/exposed domains (darker colors indicates larger weights, yellow highlight indicate if the news sources are flagged or not). In the first case, no past problematic shares were observed, but this user was heavily exposed to flagged sources over the past 2 weeks, our model allocated large attention on flagged domains from exposure (e.g., breitbart.com, thegatewaypundit.com) and on an unflagged, far-right, pro-Donald Trump cable channel (oann.com). One might suspect that the model purely generalizes past behaviors into the future. So we present the second case, both misinformation share and exposure were observed, but the model accurately predict a negative label and allocated large attention to mainstream or centered sources (washingtonpost.com, houstonchronicle.com, theguardian.com, talkingpointsmemo.com). We note that this model is unaware of any knowledge regarding a domain's learning or reliability, it learns domain-domain proximity through share/exposure sequences which in turn facilitates future prediction.

Dynamics of Misinformation Ecosystem. Besides the

temporal variation of individuals' susceptibility, we also examined the dynamics of misinformation ecosystem through a domain proximity network. Fig. 9 shows the networks for three different weeks, the first week of March, April, and May. Nodes are placed based on their proximities with every other node and simultaneously retained at similar locations across weeks. For the clarity of presentation, only 50% of nodes 624 (over 6K nodes) with the highest weighted degrees are shown. Comparing the snapshot networks, we observe that, while several most connected nodes seemed to reappear within similar clusters, there are new nodes emerging as new cluster centers as well. To see this, we list, on the top, the top 8 domains with the highest frequency appearing in users' sharing history, and also list the most connected domains having the highest degrees at the bottom for each network of the week. We found the bottom lists change rapidly, suggesting that, week by week, different flagged sites become dominant in the network as they emerged from the sequences of users' sharing. We further compute the correlation of the (weighted) degree of nodes across time, and found only 13-19% correlation (in terms of Spearman's rank correlation) between any two consecutive weeks. This result highlights the changing nature of users' sharing behavior, but at the same time demonstrate the feasibility of utilizing a deep learning framework to capture the emerging flagged domains that put new risk of misinformation to users.

Discussion

This study estimated the extent to which different user types contributed to COVID-19 misinformation spread. Our study suggests that around 90%-99% accounts that spread misinformation were humans instead of bots, over 50% bots were benign. This finding partially aligns with recent work (González-Bailón and De Domenico 2021) indicating that bots are less central than verified accounts during contentious political events, and our work provides new evidence of bot prevalence in COVID-related misinformation consumption. Regarding humans' heterogeneous misinformation behaviors, strong-adherent users exhibited more aggressive sharing activities and seemed to be more successful in spreading compared to their weak-adherent counterparts. The difference might be unsurprising as we anticipate users might have different intentions – for example, strong-adherent users might aim at supplying or/and promoting misinformation for certain purposes, whereas weak-adherent individuals might be occasionally exposed to misinformation but easily triggered to further spread misinformation. We note that, though previous work have shown concentration of volume in misinformation consumption (Grinberg et al. 2019; Guess, Nyhan, and Reifler 2018), little efforts have been made to fully understand the difference in intentions of susceptible users, which might lead to the difference in sharing tactics, temporal dedication and spread influence.

Based on our findings of different misinformation behaviors among strong-/weak-adherent users, platforms might consider devising different interventions tailored to different subgroups. For example, platforms could adopt disincentives policies (Grinberg et al. 2019) to reduce the visi-

bility of content from strong-adherent users (i.e., misinformation suppliers); on the other hand, platform could proactively deliver fact-checking corrections to weak-adherent users through their congruent friendships to boost resilience against misinformation, as our results have shown that average users have a stronger tendency to be influenced from friends, prior works also demonstrated political neutral users likely appreciate corrections of fact particularly when issued through friendships. (Parekh, Margolin, and Ruths 2020; Hannak et al. 2014; Margolin, Hannak, and Weber 2018). Future work should examine the effectiveness of such tailored intervention strategies in limiting misinformation cascades on the platform.

Our observation about political affiliation is consistent with prior work (Grinberg et al. 2019), which is expected due to the domain list we partially relied on – more political news sites among the list, and more right-leaning sites among the lower-credibility list. However, instead of focusing on political news, our study focuses on the COVID-related topics. Moreover, we found that, while CSUs are more likely than CRUs to be associated with right-leaning exposures, they tend to interact with both supporting and opposing party leaders (though in distinctive ways) as well as use both left/right-leaning hashtags in profiles. Our results suggest that, even with this global health issue, CSUs may be more responsive to politically charged COVID-related content and interactions. Our observations about emotional reactions deviates from the prior study on rumors (Vosoughi, Roy, and Aral 2018), where no substantial differences were observed in users' negative emotional reactions when interacting with false (vs. true) news. Unlike the spread of misinformation in normal times, our result suggests that during this COVID crisis, susceptible users may be more responsive to negative-emotional charged problematic content. The emotional responses were suggested by many crisis/risk literature (Inbar, Pizarro, and Bloom 2009) but we present the first large-scale empirical observation. Finally, our observation about the relationship to pre-existing susceptibility suggests the importance of looking at user susceptibility as a dynamic and situational status rather than a intrinsic trait. The proposed model could be deployed in settings where falsehood or rumors need to be paid attention early before they become entrenched in public debate, particularly during uncertain and critical times like this pandemic.

Our study has limitations. The user panel was obtained from snowball sampling and through a series of criteria, which might not be able to reflect the demographics of US population, the registered accounts on Twitter or other platforms. Another shortcoming comes from the definition of misinformation at news source level instead of at story level, along with the construction of source dictionary. Our study didn't account for the misinformation that existed outside of articles/URLs, such as false claims in text, images or videos, nor did we consider the cases that flagged sources might publish a mixture of misinformation and true stories. Besides, the major presence of conservative news sources in flagged source dictionary might partially explain the association between conservative leaning and misinformation consumption. Our findings of bot prevalence is limited by
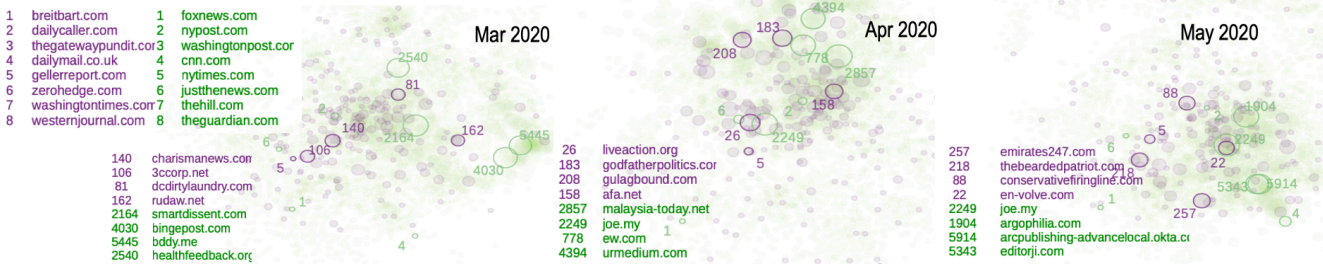
Figure 9: Domain proximity overtime. Nodes represent domains with color indicating type (purple: flagged; green: unflagged). Nodes are placed based on pairwise proximities (derived from embeddings) and simultaneously retained at similar locations across weeks. Top lists show the most connected domains with the highest frequencies to be shared over the three months, bottom lists show the most connected domains emerged from the corresponding week.

the bot detection tool. Botometer retrieved the most recent tweets to assess bot-like behaviors, the scores might be partially based on tweets posted before/after our study period. Bots are continuously evolving and novel behaviors emerge, thus our findings to some extent depend on the tool's capability in capturing such novel patterns. In addition, our analyses and prediction of linking online behavioral features to the tendency of sharing misinformation are purely correlational, more work are needed to explain why certain subgroups of people tend to share misinformation in this pandemic. Finally, though deep learning models achieved promising performance in our prediction task, they were not immune to pre-existing biases in training data, therefore they might be trained to better capture patterns from dominant subgroup of users.
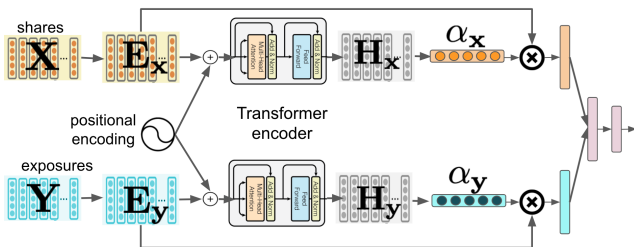
Appendix
Deep Learning Model

Figure 10: Deep learning model.

The model was inspired by prior work (Vaswani et al. 2017). In Fig. 10, both inputs – the share sequence $\mathbf{X}$ and the exposure sequence $\mathbf{Y}$ – consist of one-hot columns which could be mapped into embeddings $\mathbf{E_x}$ and $\mathbf{E_y}$, respectively. These embeddings are further encoded by a Transformer encoder into corresponding hidden states $\mathbf{H_x}$ and $\mathbf{H_y}$. The hidden states are in turn used to obtain two coefficient vectors $\alpha_\mathbf{x}$ and $\alpha_\mathbf{y}$ to indicate the contribution weights of domains from share and exposure sequences. We multiply domain embeddings with the corresponding coefficients to obtain a final representation for each input channel. Finally, a final

output layer takes the concatenation of two vectors as input and outputs a probability $\hat{y}$ to indicate the probability of future susceptibility.

As shown in Fig. 10, the output probability $\hat{y}$ can be described as below:

$$\hat{y} = \text{Softmax}(\mathbf{W}_o[\mathbf{h_x}, \mathbf{h_y}] + \mathbf{b}_o) \tag{1}$$
$$= \text{Softmax}(\mathbf{W}_o[\alpha_\mathbf{x} \odot \mathbf{W}_{emb}\mathbf{X}, \alpha_\mathbf{y} \odot \mathbf{W}_{emb}\mathbf{Y}] + \mathbf{b}_o), \tag{2}$$

where $\mathbf{W}_{emb}$ indicates embedding parameter and $\mathbf{W}_o$ ($\mathbf{b}_o$ is the weight (bias) of output layer, $\odot$ means element-wise multiplication, $[\cdot, \cdot]$ indicates the concatenation operation of two vectors. Accordingly, the contribution of domain i at time t to the final prediction is

$$CB[t, i] = \alpha_\mathbf{x}[t] \cdot \mathbf{W}_o^\mathbf{x}\mathbf{W}_{emb}[:, i], \tag{3}$$

where $\mathbf{W}_{emb}[, i]$ is domain i's embedding, and $\mathbf{W}_o^\mathbf{x}$ is the weight from $\mathbf{W}_o$ corresponding to input channel $\mathbf{X}$.

Experimental Settings

Baselines & Parameters. For our model, the feedforward dimension is 1024, the number of heads is 4, the number of encoder layers is 2. For GRU, we used bi-directional module to encode information from both channels and then concatenated the outputs for the final fully connected layer. The CNN has a combination of multiple convolution layers (with distinct filter sizes 2, 3, 4, 5) and max pooling layers. For all three neural networks, the dropout is 0.8 and the embedding dimension is 64. For the logistic regression classifier, we used the optimization solver "lbfgs."

Training Details. We prepared training data based on COVID-19 shares between March 1 and June 1, 2020. Since the panel produced a huge amount of data, we randomly sampled 30K users and we generated share/exposure sequences (observed in a 2-week period) and output labels (obtained in a 1-week period) in a sliding window manner. Long sequences of shares/exposures were truncated at maximum length 50, short sequences were abandoned at minimum length 5. After this procedure, there are 17735 unique users left and about 73000 samples as our data. The percentage of positive samples is 49%. In the training procedure, we shuffled and randomly split the samples into train 60%, validation 20% and test 20%. We ran the training procedures for 5 times and reported the average values and standard deviations for evaluation metrics.

Acknowledgments

References

Allcott, H.; and Gentzkow, M. 2017. Social Media and Fake News in the 2016 Election. *J Econ Perspect*, 31(2): 211–36.

Badawy, A.; Lerman, K.; and Ferrara, E. 2019. Who falls for online political manipulation? In *Web Conference*, 162–168.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to Ideologically Diverse News and Opinion on Facebook. *Science*, 348(6239): 1130–1132.

Ball, P.; and Maxmen, A. 2020. The Epic Battle Against Coronavirus Misinformation and Conspiracy Theories. *Nature*, 371–374.

Bessi, A.; and Ferrara, E. 2016. Social Bots Distort the 2016 US Presidential Election Online Discussion. *First Monday*, 21(11-7).

Bridgman, A.; Merkley, E.; Loewen, P. J.; Owen, T.; Ruths, D.; Teichmann, L.; and Zhilin, O. 2020. The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *HKS Misinfo. Rev.*, 1(3).

Burki, T. 2019. Vaccine Misinformation and Social Media. *Lancet Digit. Health*, 1(6): e258–e259.

Chen, E.; Lerman, K.; and Ferrara, E. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JPHS*, 6(2): e19273.

Enders, A. M.; and Smallpage, S. M. 2019. Informational Cues, Partisan-Motivated Reasoning, and the Manipulation of Conspiracy Beliefs. *Polit.Commun.*, 36(1): 83–102.

Ferrara, E. 2020. What Types of COVID-19 Conspiracies Are Populated by Twitter Bots? *First Monday*.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The Rise of Social Bots. *Commun. ACM*, 59(7): 96–104.

Freelon, D.; Marwick, A.; and Kreiss, D. 2020. False Equivalencies: Online Activism from Left to Right. *Science*, 369(6508): 1197–1201.

González-Bailón, S.; and De Domenico, M. 2021. Bots are less central than verified accounts during contentious political events. *PNAS*, 118(11).

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.

Guess, A.; Nyhan, B.; and Reifler, J. 2018. Selective Exposure to Misinformation: Evidence from the Consumption of Fake News during the 2016 US Presidential Campaign. *ERC*, 9(3): 4.

Guess, A. M.; Nyhan, B.; and Reifler, J. 2020. Exposure to Untrustworthy Websites in the 2016 US Election. *Nat. Hum. Behav*, 4(5): 472–480.

Hameleers, M.; van der Meer, T. G.; and Brosius, A. 2020. Feeling "Disinformed" Lowers Compliance with COVID-19 Guidelines: Evidence from the US, UK, Netherlands and Germany. *HKS Misinfo. Rev.*, 1(3).

Hannak, A.; Margolin, D.; Keegan, B.; and Weber, I. 2014. Get back! you don't know me like that: The social mediation of fact checking interventions in twitter conversations. In *ICWSM*.

Hua, Y.; Naaman, M.; and Ristenpart, T. 2020. Characterizing twitter users who engage in adversarial interactions against political candidates. In *CHI*, 1–13.

Hua, Y.; Ristenpart, T.; and Naaman, M. 2020. Towards measuring adversarial twitter interactions against candidates in the us midterm elections. In *ICWSM*, volume 14, 272–282.

Inbar, Y.; Pizarro, D. A.; and Bloom, P. 2009. Conservatives Are More Easily Disgusted than Liberals. *Cogn Emot.*, 23(4): 714–725.

Lazer, D.; Baum, M. A.; Ognyanova, K.; Della Volpe, J.; Quintana, A.; Chwe, H.; and McCabe, S. 2020. The state of the nation: A 50-state covid-19 survey. *The COVID-19 Consortium for Understanding the Public's Policy Preferences Across States*.

Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. 2018. The Science of Fake News. *Science*, 359(6380): 1094–1096.

Margolin, D. B.; Hannak, A.; and Weber, I. 2018. Political fact-checking on Twitter: When do corrections have an effect? *Polit Commun*, 35(2): 196–219.

Marwick, A.; and Lewis, R. 2017. Media Manipulation and Disinformation Online. *Data & Society Research Institute*.

McCorriston, J.; Jurgens, D.; and Ruths, D. 2015. Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. In *ICWSM*, 650–653. Citeseer.

Memon, S. A.; and Carley, K. M. 2020. Characterizing COVID-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791*.

Ognyanova, K.; Lazer, D.; Robertson, R. E.; and Wilson, C. 2020. Misinformation in Action: Fake News Exposure Is Linked to Lower Trust in Media, Higher Trust in Government When Your Side Is In Power. *HKS Misinfo. Rev.*

Papakyriakopoulos, O.; Serrano, J. C. M.; and Hegelich, S. 2020. The Spread of COVID-19 Conspiracy Theories on Social Media and the Effect of Content Moderation. *HKS Misinfo. Rev.*, 1(3).

Parekh, D.; Margolin, D.; and Ruths, D. 2020. Comparing Audience Appreciation to Fact-Checking Across Political Communities on Reddit. In *WebSci*, 144–154.

Roets, A.; et al. 2019. Is There An Ideological Asymmetry In the Moral Approval of Spreading Misinformation by Politicians? *Pers. Individ. Dif.*, 143: 165–169.

Roozenbeek, J.; Schneider, C. R.; Dryhurst, S.; Kerr, J.; Freeman, A. L.; Recchia, G.; Van Der Bles, A. M.; and Van Der Linden, S. 2020. Susceptibility to Misinformation about COVID-19 around the World. *R. Soc. Open Sci.*, 7(10): 201199.

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. *Nat. Commun*, 9(1): 1–9.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD*, 19(1): 22–36.

Silva, M.; Ceschin, F.; Shrestha, P.; Brant, C.; Fernandes, J.; Silva, C. S.; Grégio, A.; Oliveira, D.; and Giovanini, L. 2020. Predicting Misinformation and Engagement in COVID-19 Twitter Discourse in the First Months of the Outbreak. *arXiv preprint arXiv:2012.02164*.

Stekhoven, D. J.; and Bühlmann, P. 2012. MissForest—nonparametric Missing Value Imputation for Mixed-type Data. *Bioinformatics*, 28(1): 112–118.

Tausanovitch, C.; and Warshaw, C. 2014. Representation in municipal government. *APSR*, 605–641.

Uscinski, J. E.; Enders, A. M.; Klofstad, C.; Seelig, M.; Funchion, J.; Everett, C.; Wuchty, S.; Premaratne, K.; and Murthi, M. 2020. Why Do People Believe COVID-19 Conspiracy Theories? *HKS Misinfo. Rev.*, 1(3).

Varol, O.; Ferrara, E.; Davis, C. B.; Menczer, F.; and Flammini, A. 2017. Online Human-Bot Interactions: Detection, Estimation, and Characterization. In *ICWSM*, 280–289.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. In *NeurIPS*, 5998–6008.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The Spread of True and False News Online. *Science*, 359(6380): 1146–1151.

Zhou, X.; Zafarani, R.; Shu, K.; and Liu, H. 2019. Fake News: Fundamental Theories, Detection Strategies and Challenges. In *WSDM*, 836–837.