The Harvard Kennedy School Misinformation Review[1]
August 2020, Volume 1, Special Issue on COVID-19 and Misinformation

DOI: https://doi.org/10.37016/mr-2020-034
Website: misinforeview.hks.harvard.edu

Research Article

The spread of COVID-19 conspiracy theories on social media and the effect of content moderation

We investigate the diffusion of conspiracy theories related to the origin of COVID-19 on social media. By analyzing third-party content on four social media platforms, we show that: (a) In contrast to conventional wisdom, mainstream sources contribute overall more to conspiracy theories diffusion than alternative and other sources; and (b) Platforms' content moderation practices are able to mitigate the spread of conspiracy theories. Nevertheless, we locate issues regarding the timeliness and magnitude of content moderation, as well as that platforms filter significantly fewer conspiracy theories coming from mainstream sources. Given this, we discuss policy steps that can contribute to the containment of conspiracy theories by media sources, platform owners, and users.

Authors: Orestis Papakyriakopoulos (1), Juan Carlos Medina Serrano (1), Simon Hegelich (1)
Affiliations: (1) Political Data Science, Technical University of Munich, Germany
How to cite: Papakyriakopoulos, O.; Medina Serrano, J. C.; Hegelich, S. (2020). The spread of COVID-19 conspiracy theories on social media and the effect of content moderation. *The Harvard Kennedy School (HKS) Misinformation Review*, Volume 1, Special Issue on COVID-19 and Misinformation. https://doi.org/10.37016/mr-2020-033
Received: June 6th, 2020. Accepted: July 25th, 2020. Published: August 17th, 2020.

Research questions

- What are the dynamics of conspiracy theories related to the origin of COVID-19 on social media?
- What is the role of mainstream and alternative sources in the spread of conspiracy theories?
- What is the impact of social media platforms' content moderation policies on the diffusion of conspiracy theories?

Essay summary

- We identified 11,023 unique URLs referring to the origin of COVID-19 appearing in 267,084 Facebook, Twitter, Reddit, and 4chan posts between January and March 2020. We classified them based on their source (mainstream, alternative, other) and their content (supporting conspiracy theories, used as evidence for conspiracy theories, neither). We considered URLs in the first two content categories as stories reinforcing conspiracy theories. We investigated whether posts containing these stories were removed or labeled as such by the platforms. Then, we employed appropriate statistical techniques to quantify conspiracy theory diffusion between social media platforms and measured the impact of content moderation.

[1]*A publication of the Shorenstein Center for Media, Politics, and Public Policy, at Harvard University, John F. Kennedy School of Government.*

- We found that alternative sources generated more stories reinforcing conspiracy theories than mainstream sources. However, similar stories coming from mainstream sources reached significantly more users. We further quantified conspiracy theory dynamics in the social media ecosystem. We found that stories reinforcing conspiracy theories had a higher virality than neutral or debunking stories.
- We measured the amount of moderated content on Reddit, Twitter, and Facebook. We concluded that content moderation on each platform had a significant mitigating effect on the diffusion of conspiracy theories. Nevertheless, we found that a large number of conspiracy theories remained unmoderated. We also detected a moderation bias towards stories coming from alternative and other sources (with other sources comprising personal blogs and social media submissions, e.g. tweets, Facebook posts, Reddit comments, etc.).
- Results suggest that policymakers and platform owners should reflect on further ways that can contain COVID-19-related conspiracy theories. Content moderation is an effective strategy but can be further improved by overcoming issues of timeliness and magnitude. There should also be additional transparency on how and why content moderation takes place, as well as targeted design interventions, which can inform and sensitize users regarding conspiracy theories.

Argument & Implications

The COVID-19 health crisis resulted in the burst of an unprecedented misinfodemic on social media: A vast amount of pandemic related misinformation appeared, which in turn influenced society's response to the virus (Gyenes et al., 2018). Given the absence of exact social and scientific knowledge about the origin, nature, and impact of the coronavirus, many conspiracy theories quickly emerged, seeking to provide explanations. To confront the overwhelming amount of misinformation, social media platforms and fact-checking agencies increased attempts to moderate such content by removing or flagging it, often relying on algorithmic decision making (ADM) systems (Brennen et al., 2020; Newton, 2020).

In this study, we aim to understand how conspiracy theories spread at the beginning of the COVID-19 health crisis, and based on this, uncover possibilities and issues of fact-checking in the social media ecosystem (Marwick & Lewis, 2017). To achieve this, we measured the appearance of stories reinforcing conspiracy theories on four platforms: Facebook, Twitter, Reddit, and the subsection of 4chan called “politically incorrect” or “/pol/”, which is a prominent forum of conspiracy theorists. In contrast to other cases (Cosentino, 2020), 4chan was not the only source of conspiracy theories in the ecosystem (finding 1). We found that stories reinforcing conspiracy theories became more viral than stories either debunking them or having a neutral stance (finding 1). This complies with previous findings on misinformation (Vosoughi et al., 2018; Vicario et al., 2016).

Most of the stories reinforcing conspiracy theories originated from alternative sources, personal blogs, and social media posts (83%). However, such content coming from mainstream sources (17%) resulted in higher numbers of Facebook and Twitter shares (60% and 55% of the total respectively). Mainstream sources included high-credibility news outlets, such as the New York Post or Fox News, scientific websites such as biorxiv.org, and other widely credible sites, such as Wikipedia. Alternative sources included untrustworthy and low-credibility outlets, such as Infowars and Breitbart. Although alternative and other sources were the main carriers of conspiracy theories, mainstream sources had a higher impact on the spread of conspiracy theories (finding 2).

We investigated the platforms’ moderation practices, which varied to a certain degree. Twitter and YouTube removed stories that supported conspiracy theories (Gadde & Derella, 2020; Binder, 2020), while Reddit and Facebook either removed or flagged them (Reddit content policy, 2020; Jin, 2020). On Reddit,

removing or flagging depended on the rules of each sub-community, whereas on Facebook on whether the company reviewed the stories themselves (removed) or relied on third-party fact-checkers (flagged).

We concluded that the platforms' moderation practices strongly reduced the probability of stories reappearing in the total ecosystem. Hence, theoretically, instantly removing or filtering conspiracy theories would contain their spread. However, content moderation is a complex and time-consuming process, with human workers and ADM systems facing obstacles in accuracy and efficiency (Roberts, 2019; Graves, 2018; Gillespie, 2018; Serrano et al., 2020). This lies both in the large amount of content to be fact-checked, but also in the nature of the content, which is often difficult to categorize as conspiracy theory or not (Krafft et al. 2020; Uscinski et al. 2013; Byfold, 2011; Dentith, 2014; Krause et al., 2020). In our study, we found that the platforms managed to fact-check only between 15% to 50% of posts containing stories reinforcing conspiracy theories, with moderation in many cases taking place weeks after they became viral (finding 3,4).

We observed that each platform faced different obstacles in content moderation. For example, content moderation on Twitter was less effective than on the other platforms (finding 4). We can probably explain this effect by the timeliness of content removal, as misinformation on Twitter spreads significantly in the first hours after its first appearance (Vosoughi et al., 2018). YouTube also faced issues of timeliness. For instance, a video that stated that the pandemic is a planned conspiracy gathered up to 5 million views in a period of only two days (Wong, 2020), with copies of the video continuously being re-uploaded after its removal. Facebook filtered the least amount of stories reinforcing conspiracy theories, while Reddit appeared to not moderate older content. These results illustrate the challenges that platforms and policymakers should overcome. Besides issues of timeliness and moderation magnitude, platforms should investigate if removing or flagging content is an optimal practice, not only for containing misinformation but also for maintaining a politically inclusive environment. Since Facebook and Reddit have mixed moderation policies, it would be important to quantify different effects between misinformation control and user engagement.

A further implication of the study is related to the existence of a moderation bias on all platforms, with stories reinforcing conspiracy theories and coming from mainstream sources being filtered significantly less. This is an important finding, given that mainstream sources prevailed as a key factor for conspiracy spread in our study, and that many ADM systems for classifying contents take a source's credibility level as input (Atananosova et al., 2019). Therefore, platform owners should pay more attention to what they moderate and why, and clearly explain their decisions to the users. Studies show that additional transparency and deliberation in content removal make users more aware of the type of information they are consuming, change the way they interact with it, and build trust between them and the services (Fazio, 2020; Ruzenberg, 2019; Suzor et al., 2018; Ruzenberg, 2019; Krause et al. 2020). Finally, mainstream sources should be aware that the information they produce in the process of reportage could be exploited for the support and general reinforcement of conspiracy theories.

We hope that these recommendations can guide platforms and policymakers towards solutions that can accompany traditional content moderation, which we found to be an effective technique for containing the spread of conspiracy theories.

Findings

This study investigates content moderation practices about conspiracy theories related to the origin of COVID-19. It includes four important findings regarding conspiracy theory dynamics on social media, as well as the possibilities and issues of fact-checking for mitigating the spread of conspiracy theories.

Finding 1: URLs reinforcing conspiracy theories went more viral than URLs being neutral or debunking conspiracy theories. In both cases, URL dissemination followed complex paths in the social media ecosystem.

For the three months under investigation, we quantified the spread of conspiracy theory related URLs on social media (RQ1). Results suggest that paths and intensity varied depending on the type of URL (Figure 1). Neutral or debunking stories primarily spread in the ecosystem after being presented on Twitter, while a significant amount of URLs disseminated on other platforms through 4chan. On the other hand, stories reinforcing conspiracy theories followed different routes. URLs present on 4chan spread on Twitter, and stories on Twitter were further distributed on Facebook. Reddit had an impact on stories reinforcing conspiracy theories both on Facebook and 4chan, while Facebook was feeding 4chan with both URL types. Overall, conspiracy theory diffusion models showed that stories reinforcing conspiracy theories became more viral within the ecosystem than the rest. This complies with previous research studies stating that misinformation and provocative content is disseminated more than factual content on social networks (Vosoughi et al., 2018; Vicario et al., 2016). Furthermore, these findings show that information paths between social media are complex and content dependent, and reject the statement that fringe social media are the only contributors to conspiracy theory dissemination (Cosentino, 2020). In contrast, we found that all platforms contributed to the spread of stories reinforcing conspiracy theories.

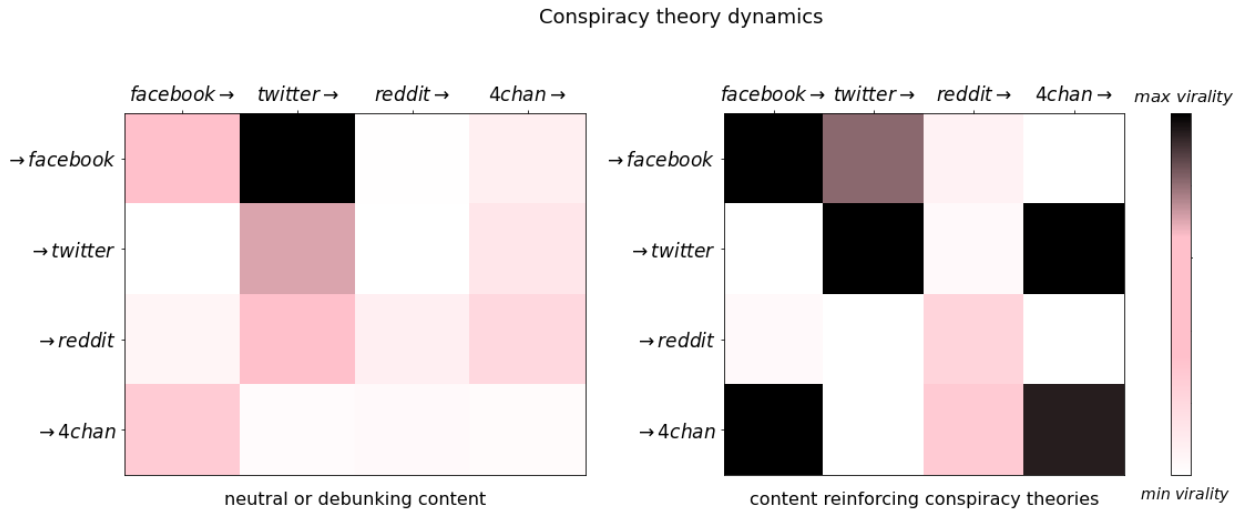

Figure 1. Heatmap depicting the virality of conspiracy theory reinforcing and neutral/debunking stories about the origin of COVID-19. *The darker the color, the higher the virality of URLs, as calculated by virality parameter α. The heatmap describes how each platform in a column affected other platforms. For example, the first column shows how stories on Facebook affected the virality of similar stories on itself, Twitter, Reddit, and 4chan.*

Finding 2: Mainstream sources played a bigger role in conspiracy theory dissemination than alternative and other sources.

We classified our sample of stories reinforcing conspiracy theories based on their source and quantified their popularity using Twitter and Facebook shares (RQ2). The 83% of the conspiracy theory reinforcing URLs originated from alternative or other sources, and only 17% came from mainstream sources. However, stories coming from mainstream sources were on average and overall more popular (Figure 2). On average, mainstream URLs supporting conspiracy theories were shared four times more on Facebook and Twitter in comparison to URLs coming from alternative sources. Similarly, mainstream URLs used as evidence for the truthfulness of conspiracy theories were shared two times more. Overall, 17% of stories

reinforcing conspiracy theories coming from mainstream sources resulted in 60% and 55% of the total Facebook and Twitter shares, respectively. These results are explainable since users usually read and share sources they trust (Brennen et al., 2020; Epstein et al., 2020), and mainstream sources have a higher reach and acceptance in society. In Table 1, we provide an exemplary set of URLs, their content type, and the number of Facebook and Twitter shares they evoked. For a more detailed analysis, refer to the appendix.

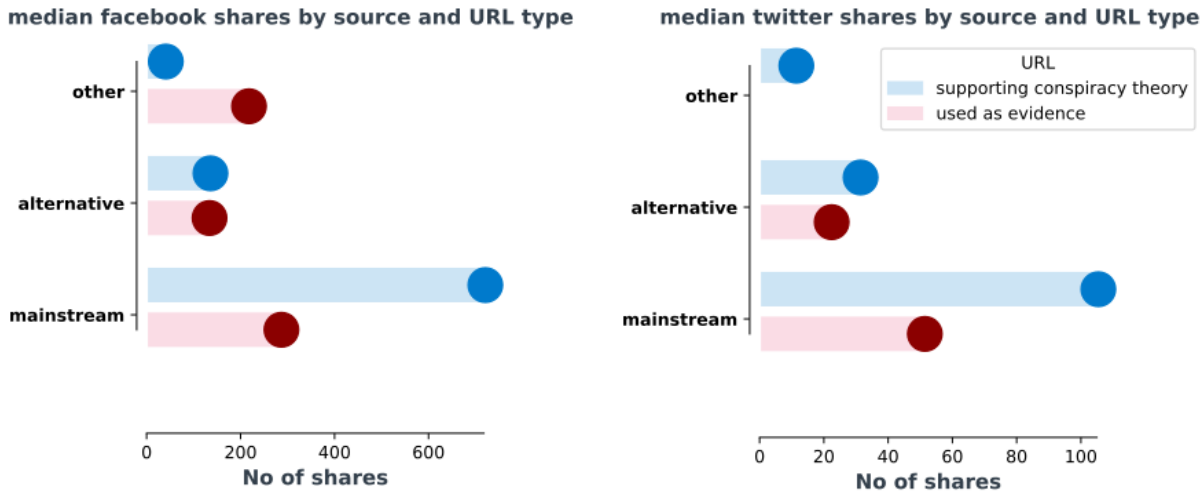

Figure 2. Bar plots illustrating median Facebook and Twitter shares for URLs supporting conspiracy theories or used as evidence by source type. *Mainstream sources included scientific articles, patent repositories, Wikipedia, government websites, high credibility and widely acceptable media outlets. Alternative sources included media outlets defined as low credibility. Other sources included social media submissions from Facebook, YouTube, Twitter, and Reddit, or personal websites and blogs.*

Table 1. *Exemplary URLs from the dataset. The table includes the title of the URL, its source, the assigned source label, content label, and the number of Facebook and Twitter shares.*

Title	Source	Source type	Content type	Facebook shares	Twitter shares
Don't buy China's story: The coronavirus may have leaked from a lab	New York Post	mainstream	supporting conspiracy theory	189,000	25,000
Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag	bioRxiv	mainstream	supporting conspiracy theory	6,500	41,000
Coronavirus Contains "HIV Insertions", Stoking Fears Over Artificially Created Bioweapon	Zerohedge	alternative	supporting conspiracy theory	14,100	15,000
Coronavirus isolated from humans	Google patents	mainstream	evidence for conspiracy theory	24,600	0

Finding 3: Moderating content either by removing or flagging it significantly reduced the spread of conspiracy theories in the ecosystem.

Information diffusion models quantified the impact of content moderation on the virality of conspiracy theories (RQ3). The models yielded for each case a value $\alpha <= 1$, which denotes the probability that a submission containing a URL will lead to the creation of another submission containing the same URL. Table 2 illustrates the mean difference of that probability when comparing models trained on URLs that

were either moderated or not. Results suggest that content moderation significantly decreased the probability that a story reinforcing conspiracy theories will reappear on the same platform, but also that it will diffuse on another platform. For Facebook and Reddit, this probability reduction exceeded 90%. By contrast, moderating content on Twitter had a smaller in-platform effect, which did not exceed 10%. A potential explanation for this is the nature of retweeting on the platform, with users spreading copies of a message in short periods after its initial submission. Thus, information can get viral before moderation mechanisms can trace it and remove it. Nonetheless, models provided evidence that content moderation practices indeed can reduce the spread of conspiracy theories in the social media ecosystem.

Table 2. *Change in the probability that a submission containing a URL will lead to the creation of another submission containing the same URL between unmoderated and moderated content. The change is given for submissions on the same platform, between platforms, and in the total ecosystem.*

	Virality reduction (%) after content moderation on:		
	Same platform	**Other platforms**	**Overall ecosystem**
Facebook	-0.96	-0.99	-0.96
Twitter	-0.10	-0.93	-0.61
Reddit	-0.97	-0.98	-0.94

Finding 4: Content moderation results revealed issues regarding the extent and nature of content removal.

Despite the finding that content moderation practices reduced the spread of conspiracy theories, our study also detected open issues when investigating RQ3. First, the biggest part of stories reinforcing conspiracy theories on the platforms remained unmoderated (between 50-85% depending on the platform) as shown on Figure 4. Especially for Reddit and Facebook, we found that if stories were not removed close to their initial submissions, the probability of them being removed later was very low. In contrast, YouTube and Twitter kept on filtering content later in time, although many of the stories had already reached peak virality. Second, we calculated the ratios of removed stories for each source type and located a source bias. On all three platforms, submissions with URLs coming from mainstream sources were removed or flagged significantly less by content moderators. This bias in content removal was translated into a relative percentage of 10 to 30 percent, depending on the platform.

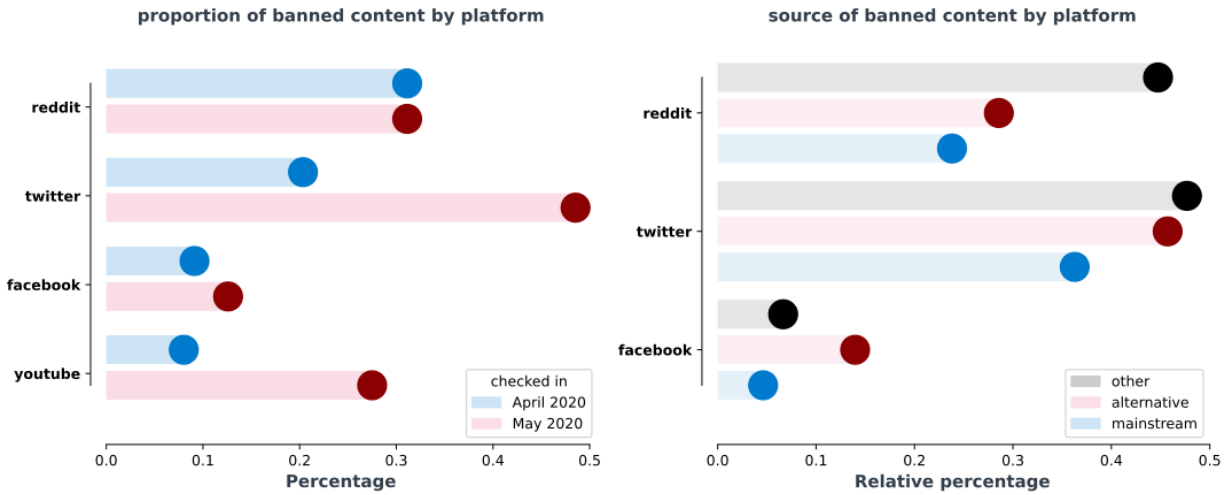

Figure 4. Left: Proportion of moderated stories reinforcing conspiracy theories as investigated in April and May 2020. Right: Relative percentage of moderated stories reinforcing conspiracy theories by source (mainstream, alternative, other).

Methods

We collected social media submissions from Reddit, Facebook, Twitter, and 4chan related to COVID-19 between January 1 and April 1, 2020. We extracted 9.5 million Reddit submissions and comments, 4.2 million Facebook posts, and 83 million tweets matching the query "COVID-19 OR coronavirus." For this, we used the Pushshift Reddit API (Baumgartner, 2018), Crowdtangle's historical data (Silverman, 2019), and the COVID-19 Twitter dataset developed by Chen et al. (2020). For 4chan, we crawled the total "Corona" thread and its sub-threads in 4chan's "politically incorrect" board and collected 1.5 million posts. From the complete dataset, we selected only the submissions that referred to the origin of COVID-19 by using the query "biowarfare OR biological weapon OR bioweapon OR umbrella corp OR man-made OR human origin OR man-made OR biosafety." We selected this query after reading multiple submissions and locating conspiracy theories that were reoccurring. As a final preprocessing step, we obtained all URLs from these submissions and created a list of 11,023 unique URLs.

We visited each of the 11,023 URLs and manually coded the stories depending on their relation to conspiracy theories. To develop a coding scheme, we adopted a definition of conspiracies and conspiracy theories based on prior theoretical work. According to Keely (1999), a conspiracy is a secret plot by two or more powerful actors. Conspiracy theories are efforts to explain the ultimate causes of significant sociopolitical events, such as the origin of COVID-19, by claiming the existence of a secret plot, by challenging institutionalized explanations (Byford, J., 2011) and many times by denying science (Douglas et al., 2019). As Byford (2011) states, conspiracy theories follow a three-point explanatory logic:

(a) **There is a conspiracy as the main narrative of a story.** For COVID-19, stories argued that a set of powerful individuals or groups, be that governments, institutions, or wealthy actors developed the virus for their specific interests.

(b) **Conspiracy theories generally ground their validity either on indirect evidence or on the absence of evidence.** For COVID-19, many stories claimed that there is a conspiracy because there exist patents on engineering coronaviruses and even a book mentioning a virus originating from Wuhan. Similarly, some stories argued that the virus should be man-made because specific research publications could not conclude on the exact animal that carried the virus.

(c) **Conspiracy theories are structured in a way that stories become irrefutable**, and hence hard to challenge (Pelkmans et al., 2011; Sunstein et al., 2009). For example, the statement "A book talked about a virus originating from Wuhan 40 years ago. Therefore, COVID-19 is man-made" is causally oversimplified and thus impossible to provide counterevidence to reject it.

By using this framework, we defined three labels for classifying URLs:

[1] **Supporting conspiracy theories.** In this case, URLs supported a conspiracy theory. The authors believed that some actors conspired to create COVID-19 and justified their thesis in the existence or absence of specific evidence.

[2] **Evidence used to support a conspiracy theory.** This class included URLs that did not directly link to a conspiracy theory, but social media users cited them as evidence for the conspiracy theories. For example, users linked to older articles about bioweapons to prove that specific countries created COVID-19. We considered this category also as reinforcing conspiracy theories because social media submissions containing these URLs were moderated by social media platforms. Furthermore, users grounded conspiracy theories on them in the way mentioned in (b).

[3] **Neither.** URLs with stories that did not refer to any type of conspiracy, that debunked conspiracy theories, mentioned conspiracy theories without believing them, or cited third parties that did believe in them.

We further labeled the URLs according to their source type. We defined three classes:

[i] **Mainstream sources.** These included scientific articles, patent repositories, Wikipedia, government websites, high credibility and widely acceptable media outlets. We used the list generated by Shao et al. (2016) and fact-checking websites (e.g. adfontesmedia.com, newsguardtech.com, allsides.com) to identify credible media outlets.

[ii] **Alternative sources.** These included media outlets defined as low credibility by Shao et al. (2016), or ranked as untrustworthy by previously mentioned fact-checking websites.

[iii] **Other sources.** These included social media submissions from Facebook, YouTube, Twitter, and Reddit, or personal websites and blogs.

To validate coding, two additional reviewers labeled a subsample of 300 URLs. The Krippendorf alpha was 0.92, while the pairwise Cohen's kappa were in all cases equal or greater than 0.9. These values suggest that there was high intercoder reliability in the labeled dataset. The subsample of 300 URLs and their corresponding reviewers' labels are available at the data repository of the study (see data availability). For further examples of our coding scheme, please refer to Table 5 in the appendix. After labeling, 4,724 URLs were supporting conspiracy theories (1) or were used as evidence of conspiracy theories (2). We searched for these URLs in the original dataset and identified 267,084 submissions that contained them.

We modeled URL cross-platform diffusion by using a mathematical technique known as Hawkes process. Hawkes process is a model that quantifies how specific events influence each other over time in an ecosystem containing multiple components. In our case, the components are the social media platforms, and an event is the appearance of a post containing a specific URL on any platform. The Hawkes process can quantify how likely it is that the posting of a URL on a platform will cause the same URL to be

posted again on any platform in the information ecosystem (Zannetou et al., 2017). This is given by a parameter α_{ij}, which gives the expected number of times a URL will appear on platform j if it was only posted on platform i, and functions as a proxy for a URL's virality (Rizoiu et al., 2017). We calculated parameters α_{ij} to study the flow of conspiracy related content across the four social media platforms, as illustrated in finding 1. For more information refer to the appendix.

To investigate the role of mainstream, alternative, and other sources in conspiracy theory dissemination, we used Buzzsumo to obtain the total number of shares each URL evoked on Facebook and Twitter. Buzzsumo provided metrics for 1,850 URLs in our sample (finding 2). We then studied whether submissions containing conspiracy theory reinforcing URLs have been removed or flagged by the platforms. We crawled Facebook, Twitter, YouTube, and Reddit once at the beginning of April 2020 and once at the beginning of May 2020 to understand content moderation (finding 4). With this new information, we ran Hawkes processes separately on moderated and unmoderated content. Finally, we compared virality parameters α for each case to quantify how much content moderation influenced conspiracy diffusion in the ecosystem (finding 3).

Bibliography

Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Karadzhov, G., Mihaylova, T., ... & Glass, J. (2019). Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality* (JDIQ), 11(3), 1-27. https://doi.org/10.1145/3297722

Baumgartner, J. M. (2018). Pushshift API.

Binder, M. (2020). YouTube tells creators to expect more video removals during coronavirus pandemic. *Mashable.* https://mashable.com/article/YouTube-coronavirus-content-moderation

Brennen, J. S., Simon, F. M., Howard, P. N., & Nielsen, R. K. (2020). Types, sources, and claims of COVID-19 misinformation. *Reuters Institute*. https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation

Byford, J. (2011). *Conspiracy theories: A critical introduction*. Springer. https://doi.org/10.1057/9780230349216

Cosentino, G. (2020). From Pizzagate to the Great Replacement: The Globalization of Conspiracy Theories. In Social Media and the Post-Truth World Order (pp. 59-86). Palgrave Pivot, Cham. https://doi.org/10.1007/978-3-030-43005-4_3

Chen, E., Lerman, K., & Ferrara, E. (2020). COVID-19: The first public coronavirus Twitter dataset. arXiv preprint arXiv:2003.07372. https://doi.org/10.2196/19273

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3), 554-559. https://doi.org/10.1073/pnas.1517441113

Dentith, M. (2014). *The philosophy of conspiracy theories*. Springer. https://doi.org/10.1057/9781137363169

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology*, 40, 3-35. https://doi.org/10.1111/pops.12568

Epstein, Z., Pennycook, G., & Rand, D. (2020, April). Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-11. https://doi.org/10.1145/3313831.3376232

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, 1(2). https://doi.org/10.37016/mr-2020-009

Gadde, V., & Derella, M. (2020). An update on our continuity strategy during COVID-19. blog. *Twitter.com*. https://blog.Twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press. https://doi.org/10.12987/9780300235029

Graves, D. (2018). Understanding the promise and limits of automated fact-checking. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves_factsheet_180226%20FINAL.pdf

Gyenes, N., & Mina, A. X. (2018). How misinfodemics spread disease. *The Atlantic*. https://www.theatlantic.com/technology/archive/2018/08/how-misinfodemics-spread-disease/568921/

Jin, K. X. (2020). Keeping people safe and informed about the coronavirus. *About Facebook*. https://about.fb.com/news/2020/07/coronavirus/

Keeley, B. L. (1999). Of conspiracy theories. *The Journal of Philosophy*, 96(3), 109-126. https://doi.org/10.2307/2564659

Krafft, P. M., & Donovan, J. (2020). Disinformation by Design: The Use of Evidence Collages and Platform Filtering in a Media Manipulation Campaign. *Political Communication*, 37(2), 194-214. https://doi.org/10.1080/10584609.2019.1686094

Krause, N. M., Freiling, I., Beets, B., & Brossard, D. (2020). Fact-checking as risk communication: the multi-layered risk of misinformation in times of COVID-19. *Journal of Risk Research*, 1-8. https://doi.org/10.1080/13669877.2020.1756385

Marwick, A., & Lewis, R. (2017). *Media manipulation and disinformation online*. New York: Data & Society Research Institute.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11), 4366-4383. https://doi.org/10.1177/1461444818773059

Newton, C. (2020, March 18). The coronavirus is forcing tech giants to make a risky bet on AI. *The Verge*. Retrieved from https://www.theverge.com/interface/2020/3/18/21183549/coronavirus-content-moderators-facebook-google-Twitter

Pelkmans, M., & Machold, R. (2011). Conspiracy theories and their truth trajectories. *Focaal*, 2011(59), 66-80. http://dx.doi.org/10.3167/fcl.2011.590105

Reddit Content policy. (2020). Reddit. https://www.redditinc.com/policies/content-policy

Rizoiu, M. A., Lee, Y., Mishra, S., & Xie, L. (2017). A tutorial on hawkes processes for events in social media. arXiv preprint arXiv:1708.06401.

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press. https://doi.org/10.2307/j.ctvhrcz0v

Roozenbeek, J., & Van Der Linden, S. (2019). The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research*, 22(5), 570-580. https://doi.org/10.1080/13669877.2018.1443491

Serrano, J. C. M., Papakyriakopoulos, O., & Hegelich, S. (2020). NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube. In *Proceedings of the ACL 2020 Workshop on Natural Language Processing for COVID-19* (NLP-COVID).

Silverman, B. (2019). CrowdTangle for academics and researchers. https://www.facebook.com/facebookmedia/blog/crowdtangle-for-academics-and-researchers

Sunstein, C. R., & Vermeule, A. (2009). Conspiracy theories: Causes and cures. *Journal of Political Philosophy*, 17(2), 202-227. https://doi.org/10.1111/j.1467-9760.2008.00325.x

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication,* 13, 18.

Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, 745-750. https://doi.org/10.1145/2872518.2890098

Uscinski, J. E., & Butler, R. W. (2013). The epistemology of fact checking. *Critical Review*, 25(2), 162-180. https://doi.org/10.1080/08913811.2013.843872

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Wong, Q. (2020). Facebook, YouTube and Twitter struggle with viral Plandemic conspiracy video. *Cnet*. Retrieved from https://www.cnet.com/news/facebook-YouTube-Twitter-viral-plandemic-conspiracy-video/

Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., Stringhini, G. & Blackburn, J. (2017). The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 Internet Measurement Conference*, 405-417.

Funding

The project was not funded by a specific organization.

Competing interests

There are no competing interests of the authors.

Ethics

The data collection and processing complied with the EU General Data Protection Regulation (GDPR).

Data availability

Based on the possibilities and limitations set by the GDPR and the guidelines of the Technical University of Munich, necessary materials and code to replicate this study are available online here: https://doi.org/10.7910/DVN/JVOUFN.

Appendix

Conspiracy theories: definition & coding

To develop a coding scheme, we adopted a definition of conspiracies and conspiracy theories based on prior theoretical work. According to Keely (1999), a conspiracy is a secret plot by two or more powerful actors. Conspiracy theories are efforts to explain the ultimate causes of significant sociopolitical events, such as the origin of COVID-19, by claiming the existence of a secret plot, by challenging institutionalized explanations (Byford, J., 2011) and many times by denying science (Douglas et al., 2019). As Byford (2011) states, conspiracy theories follow a three-point explanatory logic:

(a) **There is a conspiracy as the main narrative of a story.** For COVID-19, stories argued that a set of powerful individuals or groups, be that governments, institutions, or wealthy actors developed the virus for their specific interests.

(b) **Conspiracy theories generally ground their validity either on indirect evidence or on the absence of evidence.** For COVID-19, many stories claimed that there is a conspiracy because there exist patents on engineering coronaviruses and even a book mentioning a virus originating from Wuhan. Similarly, some stories argued that the virus should be man-made because specific research publications could not conclude on the exact animal that carried the virus. Clarke (2002) explained this type of argumentation in conspiracy theories by the fundamental attribution error bias: the tendency of humans to overstate or understate the relation between events and individuals to confirm personal dispositions.

(c) **Conspiracy theories are structured in a way that stories become irrefutable**, and hence hard to challenge (Pelkmans et al., 2011; Sunstein et al., 2009). This feature is a result of the nature of evidence used to support the conspiracy theories as mentioned in (b). For example, the statement "A book talked about a virus originating from Wuhan 40 years ago. Therefore, COVID-19 is man-made" is causally oversimplified and thus impossible to provide counterevidence to reject it.

By using this framework, we defined three labels for classifying URLs (Table 3):

[1] **Supporting conspiracy theories.** In this case, URLs supported a conspiracy theory. The authors believed that some actors conspired to create COVID-19 and justified their thesis in the existence or absence of specific evidence.

[2] **Evidence used to support a conspiracy theory.** This class included URLs that did not directly link to a conspiracy theory, but social media users cited them as evidence for the conspiracy theories. For example, users linked to older articles about bioweapons to prove that specific countries created COVID-19. They also cited a Wikipedia article about the Wuhan Biosafety lab as proof that the virus leaked from there. We considered this category as reinforcing conspiracy theories because social media submissions containing these URLs were moderated by social media platforms. Furthermore, users grounded conspiracy theories on them in the way mentioned in (b).

[3] **Neither.** URLs with stories that did not refer to any type of conspiracy, that debunked conspiracy theories, mentioned conspiracy theories without believing them, or cited third parties that did believe in them.

Table 3. *Labels used to classify the relation of a URL to conspiracy theories related to the origin of COVID-19.*

URL Label	Meaning
1	Supporting conspiracy theory
2	Evidence used to support conspiracy theory
3	Neither

We further labeled the URLs according to their source type. We defined three classes (Table 4):

[i] **Mainstream sources**. These included scientific articles, patent repositories, Wikipedia, government websites, high credibility and widely acceptable media outlets. We used the list generated by Shao et al. (2016) and fact-checking websites (e.g. adfontesmedia.com, newsguardtech.com, allsides.com) to identify credible media outlets.

[ii] **Alternative sources**. These included media outlets defined as low credibility by Shao et al. (2016), or ranked as untrustworthy by previously mentioned fact-checking websites.

[iii] **Other sources**. These included social media submissions from Facebook, YouTube, Twitter, and Reddit, or personal websites and blogs.

Table 4. *Labels used to classify the source type of the collected URLs.*

Source Label	Meaning
i	Mainstream sources
ii	Alternative sources
iii	Other sources

Table 5 presents exemplary URLs classified under each label, the reason for this, as well as an exemplary social media submission that contained them. It gives examples of URLs coming from a mainstream source, from an alternative source, as well as other sources. It also gives examples of URLs classified as supporting conspiracy theories, as evidence used to support a conspiracy theory, or neither.

Table 5. *Exemplary URLs from our dataset. The table includes the assigned source label, URL label, and the reason behind that decision. It also provides an exemplary social media submission that the URL appeared in.*

URL	Submission example	Source Label	URL Label
https://www.biorxiv.org/content/10.1101/2020.01.30.927871v1	When sections of this coronavirus directly match HIV, hard not to conclude it's an engineered bug and part of the failing NWO's plan to wrest power from populist movements the world over.	**i**	**1**
https://nypost.com/2020/02/22/dont-buy-chinas-story-the-coronavirus-may-have-leaked-from-a-lab/	The Post reported that the virus got out of a bio lab in Wuhan. If so, they may have found a way for it to avoid pre-pubescents. Crazy, scary stuff.	**i**	**1**
https://video.foxnews.com/v/6133941690001#sp=show-clips	my opinion is their is enough evidence for there being a cover up for the Wuhan bio-lab scientist releasing the man-made virus created on a joint venture between aus-USA-China after a vile was smashed during the scientists arrest in the markets, this virus 2 was designed to stop the protesting in Hong kong etc.	**i**	**1**
https://en.m.wikipedia.org/wiki/Wuhan_Institute_of_Virology	Letting the WHO genocide you with their escaped lab project would be retarded. They've been involved with the BSL4 facility from the start. They're playing dumb/savior and lying.	**i**	**2**

https://www.sciencedirect.com/science/article/pii/S0166354220300528	More evidence the virus was created in a lab to attack Wuhan. This would then be a Biowarfare DARPA PREEMPTIVE attack. The US is still the prime suspect.	**i**	**2**
https://patents.google.com/patent/US7220852B1/en	If this is true, then it could be that the UNITED STATES committed an act of BIOLOGICAL WARFARE against china!!!!	**i**	**2**
https://www.wired.com/1998/11/israels-ethnic-weapon/	Goyim, Israel has a COVID-19 vaccine, only 666 shekels, but today, just for Jew, only 660 shekels! Such a deal!"----------Israeli Scientists Claim It's 'Pure Luck' They Were Already Working On A COVID-19 Vaccine Prior To The Outbreak	**i**	**2**
https://www.zerohedge.com/economics/real-umbrella-corp-wuhan-ultra-biohazard-lab-was-studying-worlds-most-dangerous-pathogens	Alternative source for the conspiracy theorist...I do find it difficult to believe any official narrative from China!	**ii**	**1**
https://www.infowars.com/coronavirus-chinese-espionage-behind-wuhan-bioweapon/	Chinese Espionage Behind #Wuhan Bioweapon Help get the truth out by sharing this censored link!	**ii**	**1**

https://www.newstarget.com/2020-02-20-full-transcript-smoking-gun-interview-prof-frances-boyle-coronavirus-bioweapons.html	National institute of Health caught Red Handed in development and sale of Coronavirus to China!! https	**ii**	**1**
https://www.opindia.com/2020/03/korean-series-kannada-magazine-predictions-dean-koontz-sylvia-browne-chinese-coronavirus-warnings/	Someone tweaked it to increase the mortality rate to 90 per cent	**ii**	**2**
https://leozagami.com/2020/02/10/a-1981-book-by-dean-koontz-predicts-a-deadly-bacteriological-weapon-called-wuhan-400/	I did NOT write it...just seems interesting. What ya'll think? OPEN YOUR GOD DAMN EYES PEEPS 👀 DON'T BE A SHEEP	**iii**	**2**
https://www.reddit.com/r/aznidentity/comments/evxe96/novel_coronavirus_could_it_be_a_racial_bioweapon/	-	**iii**	**1**
https://www.businessinsider.com/coronavirus-white-supremacists-discussed-using-covid-19-as-bioweapon-2020-3?r=DE&IR=T	'Absolutely sickening.	**i**	**3**

https://www.foxbusiness.com/politics/coronavirus-covid-outbreak-tom-cotton	President Xi Jinping big lying about "CORONAVIRUS" outbreak...	**i**	**3**
https://www.reuters.com/article/us-health-coronavirus-disinformation/russia-deploying-coronavirus-disinformation-to-sow-panic-in-west-eu-document-says-idUSKBN21518F	Russians are not friends. They're enemies. Remember. #COVID19 #russia #TrumpPandemic	**i**	**3**
https://www.dailywire.com/news/chinese-ambassador-does-not-deny-coronavirus-came-from-biological-warfare-program	Coronavirus update: China suppressed information. Ambassador of China doesn't deny that the virus could have come from a Chinese bioweapon lab, but instead insinuates it may have come from USA weapons program.	**ii**	**3**
https://edition.cnn.com/2020/03/13/us/dean-koontz-novel-coronavirus-debunk-trnd/index.html	No, Dean Koontz did not predict the coronavirus in a 1981 novel\n#coronavirusalgerie	**i**	**3**
https://www.theepochtimes.com/involvement-of-wuhan-p4-lab-questioned_3230182.html	I think this story is more realistic about the nCoV.	**ii**	**3**

https://www.facebook.com/watch/?v=630625861056708	-	**iii**	**3**

Hawkes processes

We modeled the diffusion of conspiratorial and normal URLs in the social media ecosystem in order to understand cross-platform dynamics and the effect of content moderation practices. We assumed that in a cross-platform setting, users share contents on a platform, and other users consume it and sometimes reshare it on the same or on another platform in the ecosystem. The total life-span of a specific content in the ecosystem can be described by a point-process, i.e., a set of points in time, where each point denotes the appearance of the content. This point-process is self-exciting, meaning that the occurrence of previous points makes the occurrence of future points more probable. For example, the appearance of a tweet will trigger the appearance of a set of retweets in the future, which will not have happened without the occurrence of the initial event.

In our case, the appearance of an event (a submission containing a specific URL on a specific social media platform) can trigger the appearance of a new event on any platform in the ecosystem. A mathematical model that can describe such a multi-dimensional self-exciting point-process is the Hawkes process. A Hawkes process is a D-dimensional counting process $N(t)$=($N_1(t)\cdots N_D(t)$), where each component is a counting process:

$$N_i(t) = \sum_{k \geq 1} \mathbf{1}_{t_{i,k} \leq t}$$

, with *D* the number of social media platforms under consideration, and $t_{i,1}$,$t_{i,2}$,… being the timestamps that an event (the appearance of a specific URL) will be observed on platform i. The intensity *N* of such a process is given by the function:

$$\lambda_i(t) = \mu_i(t) + \sum_{j=1}^{D} \sum_{t_k^j < t} \phi_{ij}(t - t_k^j)$$

for i=1,…,D. Such an intensity function describes cross-platform effects induced by events created on platform *i* on events on platform j. The nature of the effects are encoded by the kernel function ϕij, t^j_k are the timestamps for all events on platform j, and μ_i is a baseline intensity that gives the magnitude of influence (how much are specific contents spread in the ecosystem in general). In our study, we use a kernel function of exponential decay, because it is able to describe content virality (how rapidly and widely is specific content diffused in the network given its previous appearance), and memory over time (for how long it remains prevalent in the ecosystem). This function is described by:

$$\phi_{ij}(t) = \alpha^{ij}\beta^{ij}\exp(-\beta^{ij}t)\,1_{t>0}$$

, where α^{ij} is the virality parameter that gives how viral content became on platform j given the appearance of content on platform i, and β^{ij} is the parameter that describes for how long contents appeared on platform j given their appearance on platform i. To calculate parameters α^{ij},β^{ij} we performed maximum likelihood estimation after splitting our data on train and test set. We fitted various Hawkes processes in order to understand differences in virality between normal and conspiratorial URLs, as well as between contents that were moderated or not for each platform.

Bibliography (Appendix)

Byford, J. (2011). *Conspiracy theories: A critical introduction*. Springer. https://doi.org/10.1057/9780230349216

Clarke, S. (2002). Conspiracy theories and conspiracy theorizing. *Philosophy of the Social Sciences*, *32*(2), 131-150. https://doi.org/10.1177/004931032002001

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. Political Psychology, 40, 3-35. https://doi.org/10.1111/pops.12568

Keeley, B. L. (1999). Of conspiracy theories. *The Journal of Philosophy*, 96(3), 109-126. https://doi.org/10.2307/2564659

Pelkmans, M., & Machold, R. (2011). Conspiracy theories and their truth trajectories. *Focaal*, 2011(59), 66-80. http://dx.doi.org/10.3167/fcl.2011.590105

Sunstein, C. R., & Vermeule, A. (2009). Conspiracy theories: Causes and cures. *Journal of Political Philosophy*, 17(2), 202-227. https://doi.org/10.1111/j.1467-9760.2008.00325.x