Online Hate Network Spreads Malicious COVID-19 Content Outside the Control of Individual Social Media Platforms

Nicolas Velasquez
George Washington University
Rhys Leahy
George Washington University
Nicholas Johnson Restrepo
George Washington University
Yonatan Lupu
George Washington University
Richard Sear
George Washington University
Nicholas Gabriel
George Washington University
Om Jha
George Washington University
Beth Goldberg
Google (United States)
Neil Johnson (neilfjohnson@me.com)
George Washington University

Research Article

Keywords: COVID-19, pandemic, clusters

Posted Date: December 3rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-110371/v1

Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms

N. Velásquez[1,2], R. Leahy[1,2], N. Johnson Restrepo[1,2], Y. Lupu[2,4], R. Sear[5], N. Gabriel[3], O. Jha[3], B. Goldberg[6], N.F. Johnson[1,2,3,*]

[1]*Institute for Data, Democracy and Politics, George Washington University, Washington D.C. 20052*
[2]*ClustrX LLC, Washington D.C.*
[3]*Physics Department, George Washington University, Washington D.C. 20052*
[4]*Department of Political Science, George Washington University, Washington D.C. 20052*
[5]*Department of Computer Science, George Washington University, Washington D.C. 20052*
[6]*Google LLC, New York City, NY 10011*

*Correspondence to neilfjohnson@me.com

Abstract: We show that malicious COVID-19 content, including racism, disinformation, and misinformation, exploits the multiverse of online hate to spread quickly beyond the control of any individual social media platform. We provide a first mapping of the online hate network across six major social media platforms. We demonstrate how malicious content can travel across this network in ways that subvert platform moderation efforts. Machine learning topic analysis shows quantitatively how online hate communities are sharpening COVID-19 as a weapon, with topics evolving rapidly and content becoming increasingly coherent. Based on mathematical modeling, we provide predictions of how changes to content moderation policies can slow the spread of malicious content.

Introduction

In addition to the spread of its biological and economic effects, the COVID-19 pandemic is playing out across the world's online platforms [1-4]. While limiting the spread of infection through social distancing, isolation has led to a surge in social media use and heightened individuals' exposure to increasingly virulent online misinformation. Users share misinformation about prevention and treatment, making it difficult for individuals to tell science from fiction. As individuals look to assign blame for the growing death toll and economic peril, extremists are rebranding their conspiracy theories around current events to draw in new followers.

This growth in hateful online activity has fueled recent attacks against vulnerable communities and government crisis responders. The AAPI reported a spike in anti-Asian assaults and harassment [5]. The FBI warned that white nationalist groups planned to spray Jews with bodily fluids from COVID-19 patients to spread the disease across the Jewish community [6]. A train operator purposefully ran a train off the rails at full speed near the USNS *Mercy*, a coronavirus relief ship, because he believed the effort was covering up a government takeover [7].

Winning the war against such malicious online content will require an understanding of the entire online battlefield. A rich literature across many disciplines explores the problem of online misinformation [8-12], detailing some suggestions for how social media platforms can address the problem [1, 13-17]. However, the bulk of existing work focuses on the spread of

misinformation *within a single* platform, e.g. Twitter, but contemporary social media platforms are not walled gardens.

As we show in this paper, combating online misinformation requires an analysis of how it spreads *across multiple* social media platforms. Each social media platform is effectively its own *universe*, i.e., a commercially independent entity subject to particular legal jurisdictions [21,22], but these universes are connected to each other by users and their communities. We show that hate communities spread malicious COVID-19 content across social media platforms in ways that subvert the moderation attempts of individual platforms. Moreover, there is now a proliferation of other, far less regulated platforms thanks to open-source software enabling decentralized setups across locations. Cooperation by moderators across platforms is unlikely because of competing commercial incentives -- therefore we develop implications for policing approaches to reduce the diffusion of malicious online content that do not rely on future global collaboration across social media platforms.

Design and Results

To gain a better understanding of how malicious content spreads, we begin by creating a map of the network of online hate communities across six social media platforms. We include mainstream platforms -- Facebook, VKontakte, and Instagram -- that have and enforce (to varying degrees) policies against hate speech, as well as fringe platforms with minimal content policies: Gab, Telegram, and 4Chan. Each of these platforms allows users to create and join interest-based communities, (e.g., Facebook page, VKontakte group, Telegram channel,) which we refer to as "clusters." Within such clusters, users develop and coordinate around narratives -- in contrast to platforms like Twitter that have no in-built community tool and are instead designed for broadcasting short messages [18-20].

We include in our data, clusters in which 2 out of the 20 most recent posts at the time of classification include hate content. We define hate content as either (a) content that would fall under the provisions of the United States' Code regarding hate crimes or hate speech according to Department of Justice's guidelines, or (b) content that supports or promotes fascist ideologies or regime types (e.g., extreme nationalism and/or racial identitarianism).

Using this methodology, we identified approximately 6000 online hate clusters across these platforms, involving approximately 10 million users across the globe. Examples of clusters include a Facebook fan page in which the Russian Imperial Movement -- a designated terrorist organization -- promotes its training camp, a Telegram channel in which former Wisconsin Republican primary candidate Paul Nehlen advocates murdering Jews, and an Australian Gab group in which members fantasize about founding a white-only colony in space. Most analyses of online extremist activity focus on a single platform, but extremists, like anyone else, simultaneously use multiple platforms for complementary functions. This redundancy helps extremist networks develop resilience across the multiverse of platforms. An extremist group might maintain a Facebook page, Instagram, or Twitter account where they share incendiary news stories and spicy memes to draw in new followers. These accounts might walk right up to the line dividing hate speech from political speech, but they won't cross it. Once they've built interest and gained the trust of those new followers, the most active

members and page administrators route other people to less moderated platforms like VKontakte or Gab where they can openly discuss hateful and extreme ideologies.

Next, we identified and mapped out online hyperlinks across clusters and across platforms using the methodology described in [18,19] (see also Methods and SI). Then we identified malicious content related to COVID-19 by searching for usage of specific keywords and constructs related to the pandemic. This terminology differs by time period given the quickly evolving nature of the pandemic. For example, terms such as "COVID-19" and "SARS-CoV-2" were officially introduced by the World Health Organization in February 2020, prior to when it was colloquially known in the hate clusters by names such as "Chinese Zombie Virus", and "Wuhan Virus." The SI provides details and examples of this material.

To understand more fully the dynamics by which COVID-19 content diffuses and evolves across the online hate network, and to inform the policy solutions offered in Fig. 3, we conduct three analyses. First, we analyze the connectivity of clusters across the moderated and unmoderated platforms. Figure 1 shows the percentage of links within the hate network that are between given pairs of platforms. For example, 61.95% of the cross-platform links in the network are from VKontakte into Telegram, while 15.22% of the links are from Gab into 4chan. In part because of content moderation, only two platforms connect outward to all the other platforms: Telegram and Vkontakte. If a Facebook user posts a link to Gab, for example, such a link would be removed by content moderators. However, a Facebook user can link to a VKontakte page that links to the same Gab cluster, and these indirect links allow users to access misinformation and hate content across the multiverse of platforms.

This means that links across social media platforms act like wormholes to create a huge, decentralized multiverse that connects hate communities. We therefore analyze in more detail how malicious COVID-19 content diffuses across the hate network. Figures 2A and B describe the map of this online hate network.

Each of the hate clusters appears as a node with a black circle, while other clusters linked to by hate clusters appear as nodes without black circles. Figures 2A and B show how COVID-19 malicious content is exploiting the existing online hate network to spread quickly between platforms and hence beyond the control of any single platform.

We then analyze how malicious COVID-19 content evolves in the hate network (following the methodology described in [23]). We conduct machine-learning topic analysis using Latent Dirichlet Allocation (LDA) [24] to analyze the emergence and evolution of topics around COVID-19. We then calculate a coherence score, which provides a quantitative method for measuring the alignment of the words within an identified topic [24]. Figure 3 provides an example of the results of this analysis within a single hate cluster. We find that the coherence of COVID-19 discussion increased rapidly in the early phases of the pandemic, with narratives forming and cohering around COVID-19 topics and misinformation.

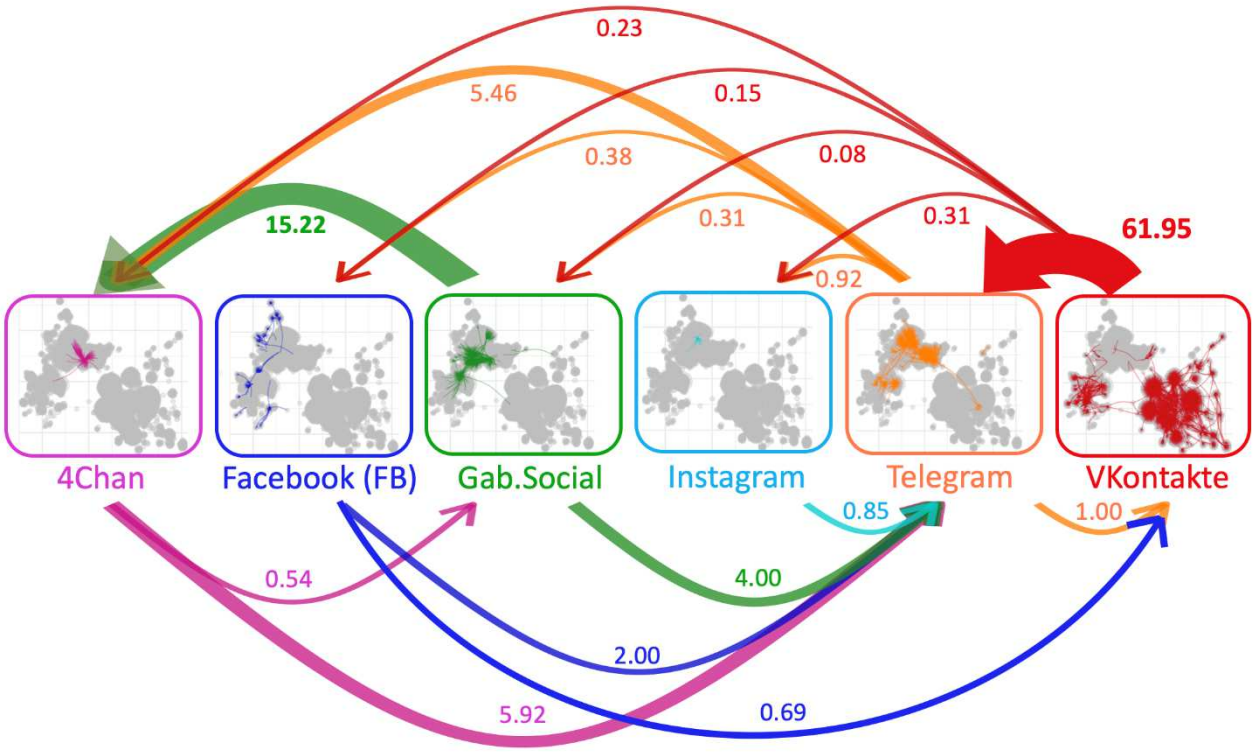

Figure 1: Connectivity across the Online Hate Multiverse. We counted all links between hate clusters on different social media platforms between June 1, 2019 and February 1, 2020. Each arrow shows the percentage of such links from hate clusters on the outbound platform to hate clusters on the inbound platform. Some platform pairs feature either zero such links or a negligible amount, hence an arrow is not shown. Although content moderation prevents users on some platforms (e.g., Facebook) from linking to some unmoderated platforms (e.g., Gab), users can access such content – and direct other users to it – by linking to a hate cluster on a third platform (e.g., VKontakte) that, in turn, links to the unmoderated platform.

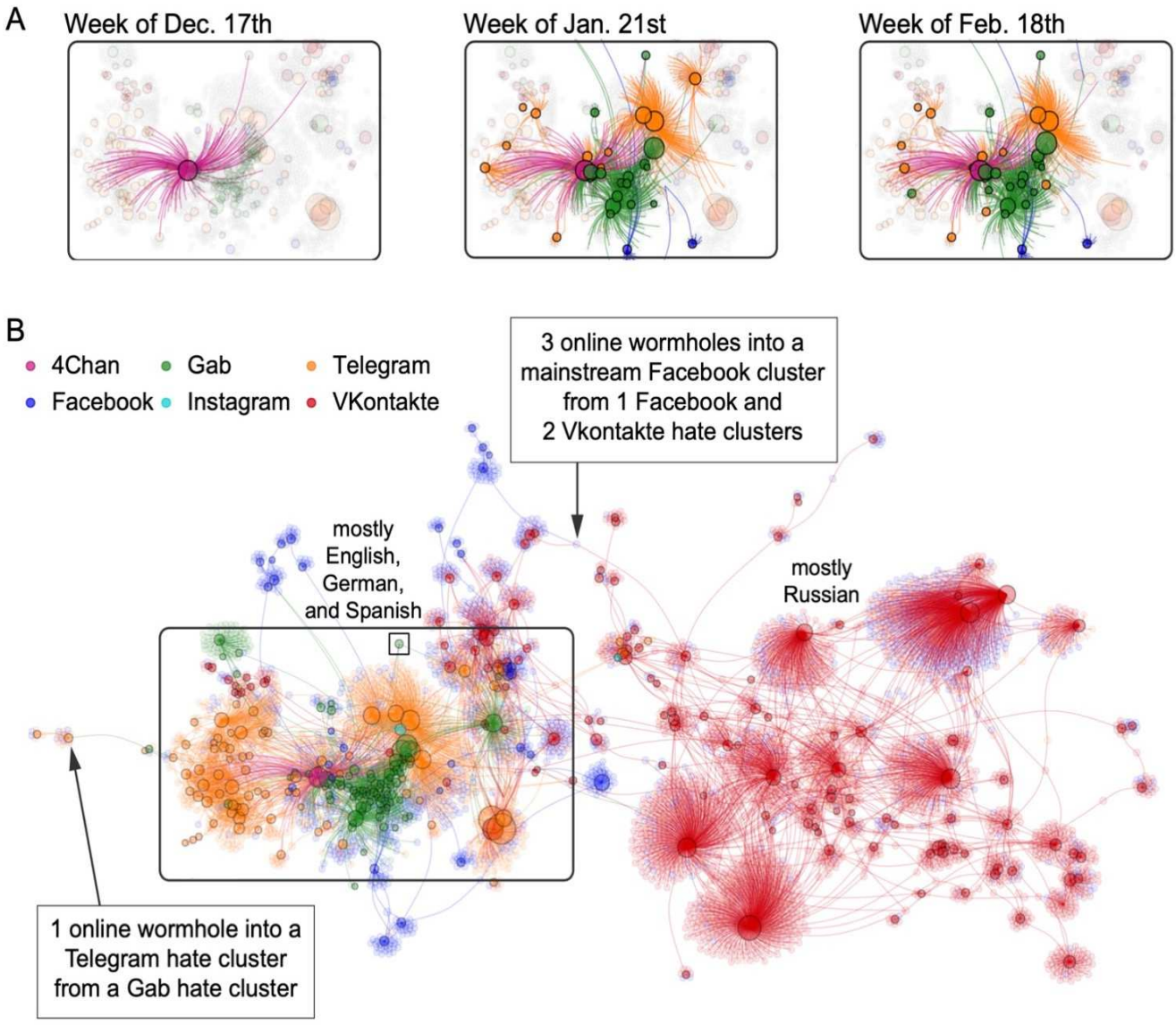

Figure 2: Malicious COVID-19 content spreading across the Online Hate Multiverse. A: Time evolution of birth and spread of malicious COVID-19 content within and across different social media platforms within a portion of the online hate network in B outlined in black. B: The online hate multiverse comprises separate social media platforms that interconnect over time via dynamic connections created by hyperlinks from clusters on one platform into clusters on another. Links shown are from hate clusters (i.e., online communities with hateful content, shown as nodes with black rings) to all other clusters, including mainstream ones (e.g., football fan club). Link color denotes platform hosting the hate cluster from which link originates. Plot aggregates activity from June 1st, 2019 to March 23rd, 2020 and should be similar at time of publication. The observed layout is spontaneous (i.e., not built-in, see Methods). The small black square (inside the larger black square) is the Gab cluster analyzed in Fig. 3 (see Methods and SI for details).

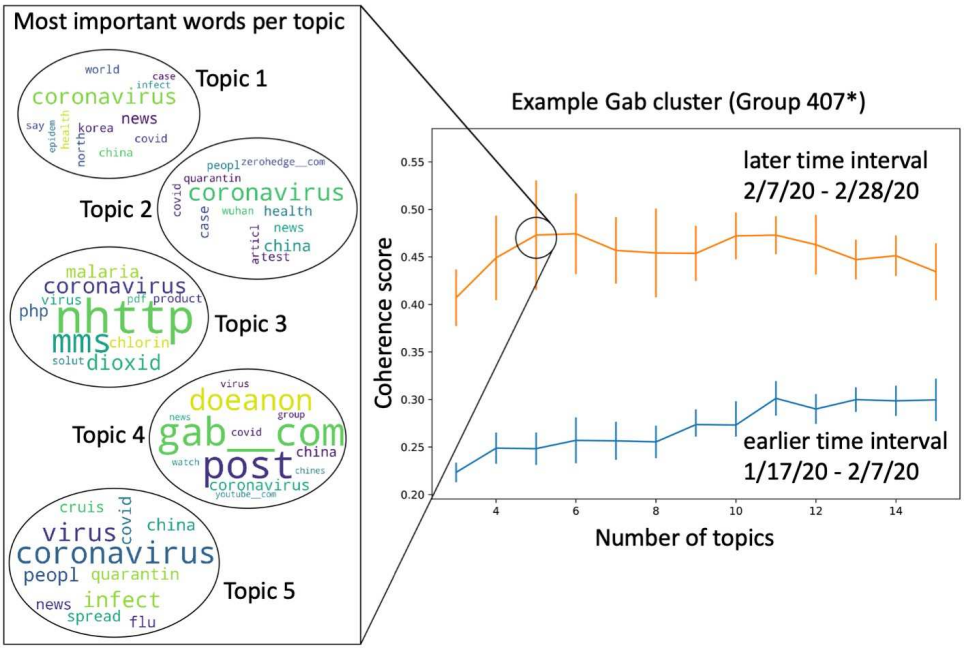

Figure 3: Evolution of COVID-19 content. Focusing on a single Gab hate cluster, this provides example output from our machine learning topic analysis of content. Even though discussion of COVID-19 only arose in December 2019, it quickly evolved from featuring a large number of topics with a relatively low average coherence score, to featuring a small number of topics with high average coherence score more focused around COVID-19. As the right-hand panel shows, the discussion of in this cluster became much more coherent, and focused on COVID-19, during the second three-week period we analyzed. The right-hand panel shows the keywords in each of 5 topics discussed on this cluster during that second three-week period. In the first three-week period, topics discussed featured profanity and hate speech such as f* and n*****, but the conversation quickly become more focused and less like a stereotypical hate-speech rant. SI shows explicit examples of this content.**

Discussion

The first general implication of our findings is that in order to understand the diffusion of COVID-19 and related malicious matter, we need to account for the decentralized, interconnected nature of this online network (Fig. 2). Links connecting clusters on different social media platforms provide a gateway that can pass malicious content (and supporters) from a cluster on one platform to a cluster on another platform that may be very distant geographically, linguistically, and culturally, e.g., from Facebook to VKontakte. Figure 2 shows that consecutive use of these links allows malicious matter to find short pathways that cross the entire multiverse, just as short planks of wood can be used to bridge adjacent rocks and cross a wide river. Because malicious matter frequently carries quotes and imagery from different moments in a cluster's timeline, these inter-platform links not only interconnect information from disparate points in space, but also time -- like a wormhole.

A second implication is that malicious activity can appear isolated and largely eradicated on a given platform, when in reality it has moved to another platform. There, malicious content can thrive beyond the original platform's control, be further honed, and later *reintroduced into the original platform* using a link in the reverse direction. Facebook content moderators reviewing only Facebook (i.e., blue) clusters in Fig. 2B might conclude that they had largely rid that platform of hate and disconnected hateful pages from one another, when in fact these same clusters remain connected via other platforms. Because the number of independent social media platforms is growing, this multiverse will continue to grow and will likely be fully connected via new links.

Implication 3 is that this multiverse acts like a global funnel that can suck individuals from a mainstream cluster on a platform that invests significant resources in moderation, into less moderated platforms like 4Chan or Telegram, simply by offering them wormhole links to follow. As Fig. 1 illustrates, an innocent user of mainstream social media communities, including a child connecting with other online game players or a parent seeking information about COVID-19, is at most a few links away from intensely hateful content. In this way, the rise of fear and misinformation around COVID-19 has allowed promoters of malicious matter and hate to engage with mainstream audiences around a common topic of interest, and potentially push them toward hateful views.

Implication 4 is that it is highly unlikely that the multiverse in Fig. 2B is, or could be, controlled by a single state actor, given its vast decentralized nature. We have checked, for example, for evidence of explicit Russian-sponsored campaigns. Because many hate clusters organize around the topics of minorities and refugees, we expected to find frequent links to Russian media, but instead only found a small portion of clusters linking to Kremlin-affiliated domains. These links accounted for $<0.5\%$ of all posts shared. This is also consistent with the notion that the extended nature of exchanges in a cluster enables a community to collectively weed out coordinated trolls and bot-like members.

Implication 5 comes from the topic analysis described in Figure 3. This shows that the discussion within the global online hate community has coalesced around COVID-19, with topics evolving rapidly and their coherence scores increasing. Examples of weaponized content (see SI) reveal evolving narratives such as blaming Jews and immigrants for inventing and spreading the virus, and instances of neo-Nazis planning attacks on emergency responders to the health crisis. While these topics morph, the underlying structure in Fig. 2B remains rather robust, which suggests that our implications should also hold in the future.

In summary, no single platform can address the problem of malicious COVID-19 content, yet coordinated moderation among all platforms (some of which are unmoderated) is highly unlikely. We therefore offer predictions based on a mathematical model that suggest that platforms could use bilateral link engineering to artificially lengthen the pathways that malicious matter needs to take between clusters, increasing the chances of its detection by moderators and delaying the spread of time-sensitive material such as weaponized COVID-9 misinformation and violent content (see SI for details).

This involves the following repeated process: first, pairs of platforms use Fig. 2B to estimate the likely numbers of wormholes or indirect ties between them. Then, without having to exchange any sensitive data, each can use our mathematical formulae (see SI) to engineer the correct cost w for malicious content spreaders who are exploiting their platform as a pathway, i.e.., they can focus available moderator time to achieve a particular detection rate for malicious material passing through their platform and create an effective cost w for these spreaders in terms of detection, shut-down, and sanction. While Figs. 4A and 4B show common situations that arise in Fig. 2B, more complex combinations can be described using similar calculations (see SI) in order to predict how the path lengths for hate material can be artificially extended in a similar way to Fig. 4C.

Our predictions (see SI) show that an alternative though far more challenging way of reducing the spread of malicious content is by manipulating either (1) the size N of its online potential supporters (e.g., by placing a cap on the size of clusters) and/or (2) their heterogeneity F (e.g., by introducing other content that effectively dilutes a cluster's focus). Figure 4D shows examples of the resulting time-evolution of the online support, given by $N\left(1 - W\left(\left[\frac{-2Ft}{N}\right] exp\left[\frac{-2Ft}{N}\right]\right) / \left[\frac{-2Ft}{N}\right]\right)$ where the resulting delayed onset time for the rise in support is $t_{onset} = \frac{N}{2F}$ and where W is the Lambert function [28]. Figures 4E and F show related empirical findings which are remarkably similar to Fig. 4D. Figure 4F is a proxy system [27] in which ultrafast predatory algorithms began operating across electronic platforms to attack a financial market order book in subsecond time [27]. Figure 4F therefore also serves to show what might happen in the future if the hate multiverse in Fig. 2B were to become populated by such predatory algorithms whose purpose is now to quickly spread malicious matter. Worryingly, Fig. 4F shows that this could result in a multiverse-wide rise in malicious matter on an ultrafast timescale that lies beyond human reaction times [27].

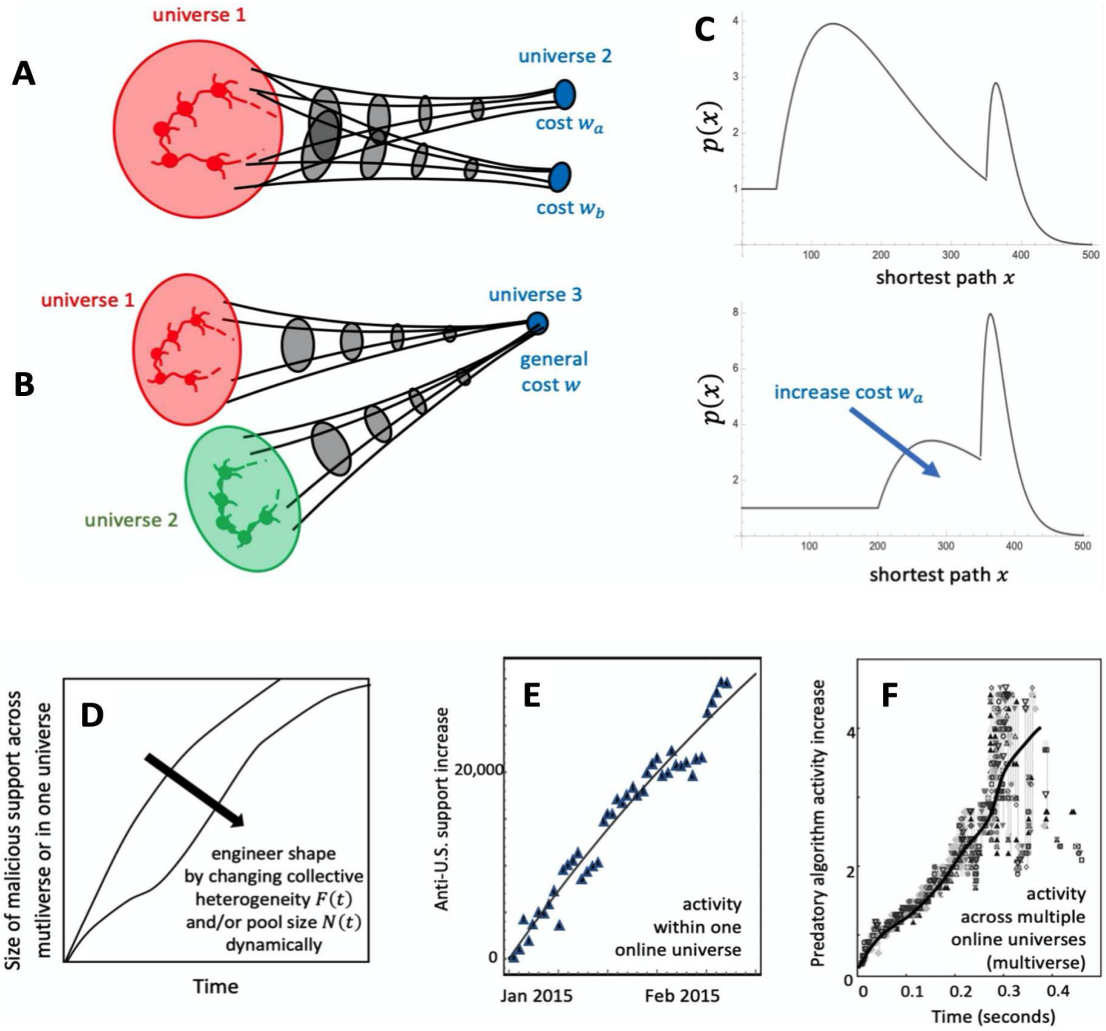

Figure 4: Wormhole Engineering to Mitigate Spreading. A-B: Typical motifs within the full multiverse in Fig. 2B. C: Mathematical prediction for motif A, showing that the distribution of shortest paths (top panel, shown un-normalized) for transporting malicious matter across a platform (i.e., universe 1) can be shifted to larger values (bottom panel) which will then delay spreading and will increase the chance that the malicious matter is detected and removed [25,26]. This is achieved by manipulating the risk that the malicious content gets detected when passing via the other platform: this risk represents a cost for the hate community in universe 1 when using the blue node(s). The same mathematics applies irrespective of whether each blue node is a single cluster or an entire platform, and applies when both blue clusters are in the same platform or are in different platforms. See SI for case B.
D-F: Mathematical prediction that the total online support for malicious matter can be manipulated by varying the online pool size of potential supporters $N(t)$ and/or their heterogeneity $F(t)$. The mathematics we develop here has implications beyond the hate network shown in Fig. 2B. E: Example of how an empirical outbreak of anti-U.S. hate across a single platform (VKontakte) in 2015 produces similar shape to upper curve in D. F: Empirical outbreak for the proxy system of predatory 'buy' algorithms across multiple electronic platforms [27] also produces a similar shape to lower curve in D. (See SI for details).

This analysis of course requires follow-up work. Our mathematical formulae are, like any model, imperfect approximations. However, we have checked that they agree with large-scale numerical simulations [25-29] and follow similar thinking to other key models in the literature [30-32]. Going forward, other forms of malicious matter and messaging platforms need to be included. However, our initial analysis suggests similar findings for any platforms that allow communities to form. We should also further our analysis of the time-evolution of cluster content using the machine-learning topic modeling approach and other methods. We could also define links differently, e.g., numbers of members that clusters have in common. However, such information is not publicly available for some platforms, e.g., Facebook. Moreover, our prior study of a Facebook-like platform where such information was available showed low/high numbers of common members reflects the absence/existence of a cluster-level link, hence these quantities indeed behave similarly to each other. People can be members of multiple clusters; however, our prior analyses suggest only a small percentage are *active* members of multiple clusters. In terms of how people react to intervention, it is known that some may avoid opposing views [33] while for others it may harden beliefs [34]. However, what will actually happen in practice remains an empirical question.

Methods

Humans are not directly involved in this study. Our methodology focuses on aggregate data about online clusters and posts, hence the only data required that involves individuals is the open source content of their public posts, which is publicly available information -- just as information about a specific molecule of water is not needed to describe the bubbles (i.e., clusters of correlated molecules) that form in boiling water. Links between clusters are hyperlinks. Our network analysis for Fig. 2B starts from a given hate cluster A and captures any cluster B to which hate cluster A has shared an explicit cluster-level link. We developed software to perform this process automatically and, upon cross-checking the findings with our manual list, were able to obtain approximately 90 percent consistency between manual and automated versions. All but one node in Fig. 2B is plotted using the ForceAtlas2 algorithm, which simulates a physical system where nodes (clusters) repel each other while links act as springs, and nodes that are connected through a link attract each other. Hence nodes (clusters) closer to each other have more highly interconnected local environments while those farther apart do not. The exception to this Force Atlas2 layout in Fig. 2B is Gab group 407* ("Chinese Coronavirus", https://gab.com/groups/407*, see small black square in Fig. 2B) which was manually placed in a less crowded area to facilitate its visibility. This particular cluster was created in early 2020 with a focus on discussing the COVID19 pandemic, but it immediately mixed hate with fake news and science, as well as conspiratorial content.

Data Availability: Humans are not directly involved in this study. Aggregate information data will be provided with the Supplementary Information (SI). All computer programs are described fully in previous publications (see references).

References

1. Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, *31*(7), 770-780.

2. Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., ... & Drury, J. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, 1-12.

3. Brennen, J. S., Simon, F., Howard, P. N., & Nielsen, R. K. (2020). Types, sources, and claims of Covid-19 misinformation. *Reuters Institute*, *7*, 3-1.

4. Cuan-Baltazar, J. Y., Muñoz-Perez, M. J., Robledo-Vega, C., Pérez-Zepeda, M. F., & Soto-Vega, E. (2020). Misinformation of COVID-19 on the internet: infodemiology study. *JMIR public health and surveillance*, *6*(2), e18444.

5. Reports of Anti-Asian Assaults, Harassment and Hate Crimes Rise as Coronavirus Spreads, Anti-Defamation League, April 8, 2020. https://www.adl.org/blog/reports-of-anti-asian-assaults-harassment-and-hate-crimes-rise-as-coronavirus-spreads.

6. Far-right and radical Islamist groups are exploiting coronavirus turmoil, The Washington Post, April 10, 2020. https://www.washingtonpost.com/national-security/far-right-wing-and-radical-islamist-groups-are-exploiting-coronavirus-turmoil/2020/04/10/0ae0494e-79c7-11ea-9bee-c5bf9d2e3288_story.html.

7. Train Operator at Port of Los Angeles Charged with Derailing Locomotive Near U.S. Navy's Hospital Ship Mercy, Department of Justice, U.S. Attorney's Office, Central District of California, April 1, 2020. https://www.justice.gov/usao-cdca/pr/train-operator-port-los-angeles-charged-derailing-locomotive-near-us-navy-s-hospital.

8. Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016, April). Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web* (pp. 745-750).

9. Ratkiewicz J, Conover M, Meiss M, Goncalves B, Flammini A, Menczer F. Detecting and Tracking Political Abuse in Social Media. In: Proc. International AAAI Conference on Web and Social Media. Palo Alto, CA: AAAI; 2011. p. 297–304. Available from: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2850.

10. C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In Proceedings of the 20th international conference on World wide web, pages 675--684. ACM, 2011

11. Sampson J, Morstatter F, Wu L, Liu H. Leveraging the Implicit Structure Within Social Media for Emergent Rumor Detection. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. CIKM'16. New York, NY, USA: ACM; 2016. p. 2377–2382. Available from: http://doi.acm.org/10.1145/2983323.2983697.

12. Ferrara E. Disinformation and social bot operations in the run up to the 2017 French presidential election. First Monday. 2017;22(8).

13. Wardle C. Fake news. It's complicated. First Draft News; 2017. Available from: https://firstdraftnews.com/fake-news-complicated/.

14. Nguyen, N. P., Yan, G., Thai, M. T., & Eidenbenz, S. (2012, June). Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference* (pp. 213-222).

15. He, Z., Cai, Z., Yu, J., Wang, X., Sun, Y., & Li, Y. (2016). Cost-efficient strategies for restraining rumor spreading in mobile social networks. *IEEE Transactions on Vehicular Technology*, *66*(3), 2789-2800.
Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.

16. Chou, W. Y. S., Oh, A., & Klein, W. M. (2018). Addressing health-related misinformation on social media. *Jama*, *320*(23), 2417-2418.

17. Pennycook, G., Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521-252

18. Johnson, N.F., R. Leahy, N. Johnson Restrepo, N. Velasquez, M. Zheng, P. Manrique, P. and S. Wuchty. Hidden resilience and adaptive dynamics of the global online hate ecology. Nature 573, 261 (2019)

19. Johnson, N.F., M. Zheng, Y. Vorobyeva, A. Gabriel, H. Qi, N. Velasquez, P. Manrique, D. Johnson, E. Restrepo, C. Song, S. Wuchty. New online ecology of adversarial aggregates: ISIS and beyond. Science 352, 1459 (2016)

20. Ammari, T., S. Schoenebeck. "Thanks for your interest in our Facebook group, but it's only for dads:" Social Roles of Stay-at-Home Dads. CSCW '16, February 27-March 02, 2016, San Francisco, CA, USA. http://dx.doi.org/10.1145/2818048.2819927

21. Iyengar, R. The coronavirus is stretching Facebook to its limits. March 18, 2020. See https://www.cnn.com/2020/03/18/tech/zuckerberg-facebook-coronavirus-response/index.html

22. Frenkel, S., D. Alba and R. Zhong. Surge of Virus Misinformation Stumps Facebook and Twitter. The New York Times, March 8, 2020: www.nytimes.com/2020/03/08/technology/coronavirus-misinformation-social-media.html

23. Sear, R.F., Velasquez, N., Leahy, R., Restrepo, N. J., El Oud, S., Gabriel, N., Lupu, Y. & Johnson, N. F. (2020). Quantifying COVID-19 content in the online health opinion war using machine learning. *IEEE Access*.

24. Syed, S., M. Spruit, ‘‘Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation,’’ in Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA), Oct. 2017, pp. 165–174, doi: 10.1109/DSAA.2017.61.

25. Ashton, D.J., T.C. Jarrett and N.F. Johnson. Effect of Congestion Costs on Shortest Paths Through Complex Networks. Phys. Rev. Lett. 94, 058701 (2005).

26. Jarrett, T.C., D.J. Ashton, M. Fricker and N.F. Johnson. Interplay between function and structure in complex networks. Phys. Rev. E 74, 026116 (2006).

27. Johnson, N.F. To slow or not? Challenges in subsecond networks. Science 355, 801 (2017).

28. Manrique, P.D., M.Zheng, Z. Cao, E.M. Restrepo, N.F. Johnson . Generalized gelation theory describes onset of online extremist support. Phys. Rev. Lett. 121, 048301 (2018).

29. Zhao, Z., J.P. Calderon, C. Xu, G. Zhao, D. Fenn, D. Sornette, R. Crane, P.M. Hui, N.F. Johnson. Effect of social group dynamics on contagion. Phys. Rev. E 81, 056107 (2010)

30. Gavrilets, S. Collective action and the collaborative brain. J. R. Soc. Interface 12, 20141067 (2015)

31. Havlin, S., D.Y. Kenett, A. Bashan, J. Gao, H.E. Stanley. Vulnerability of network of networks. The European Physical Journal Special Topics. 223, 2087 (2014)

32. Palla, G., A.L. Barabasi, T. Vicsek. Quantifying social group evolution. Nature 446, 664 (2007)

33. Frimer, J.A., L.J. Skitka, M. Motyl. Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. Journal of Experimental Social Psychology, 72 (2017) 10.1016/j.jesp.2017.04.003

34. Bail, C.A., L.P. Argyle, T.W. Brown, J.P. Bumpus, H. Chen, M.B. Fallin Hunzaker, J. Lee, M. Mann, F. Merhout, A. Volfovsky. Exposure to opposing views on social media can increase political polarization. PNAS 115, 9216-9221 (2018) https://doi.org/10.1073/pnas.1804840115

Figures

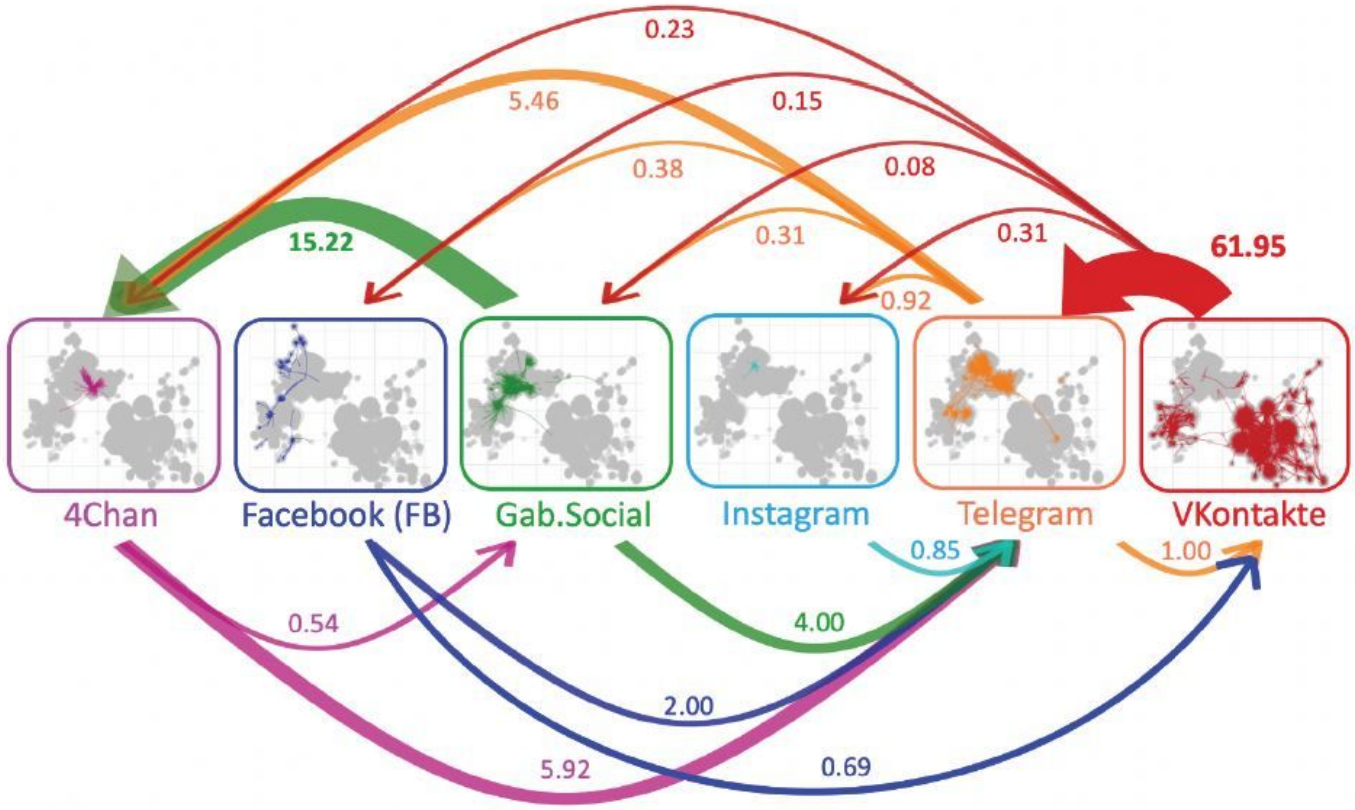

Figure 1

Connectivity across the Online Hate Multiverse. We counted all links between hate clusters on different social media platforms between June 1, 2019 and February 1, 2020. Each arrow shows the percentage of such links from hate clusters on the outbound platform to hate clusters on the inbound platform. Some platform pairs feature either zero such links or a negligible amount, hence an arrow is not shown. Although content moderation prevents users on some platforms (e.g., Facebook) from linking to some unmoderated platforms (e.g., Gab), users can access such content – and direct other users to it – by linking to a hate cluster on a third platform (e.g., VKontakte) that, in turn, links to the unmoderated platform.

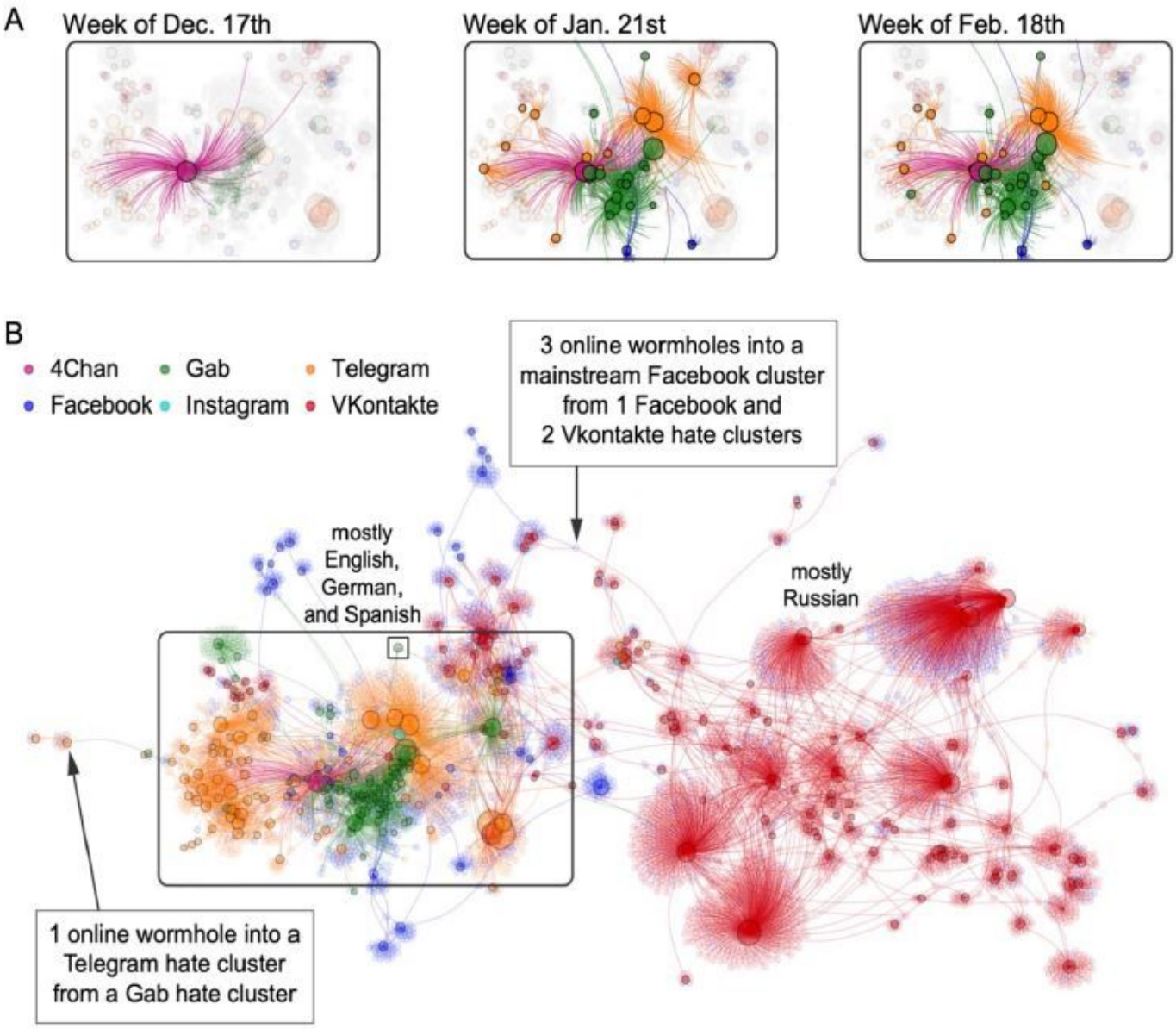

Figure 2

Malicious COVID-19 content spreading across the Online Hate Multiverse. A: Time evolution of birth and spread of malicious COVID-19 content within and across different social media platforms within a portion of the online hate network in B outlined in black. B: The online hate multiverse comprises separate social media platforms that interconnect over time via dynamic connections created by hyperlinks from clusters on one platform into clusters on another. Links shown are from hate clusters (i.e., online communities with hateful content, shown as nodes with black rings) to all other clusters, including mainstream ones (e.g., football fan club). Link color denotes platform hosting the hate cluster from which link originates. Plot aggregates activity from June 1st, 2019 to March 23rd, 2020 and should be similar at time of publication. The observed layout is spontaneous (i.e., not built-in, see Methods). The small black square (inside the larger black square) is the Gab cluster analyzed in Fig. 3 (see Methods and SI for details).

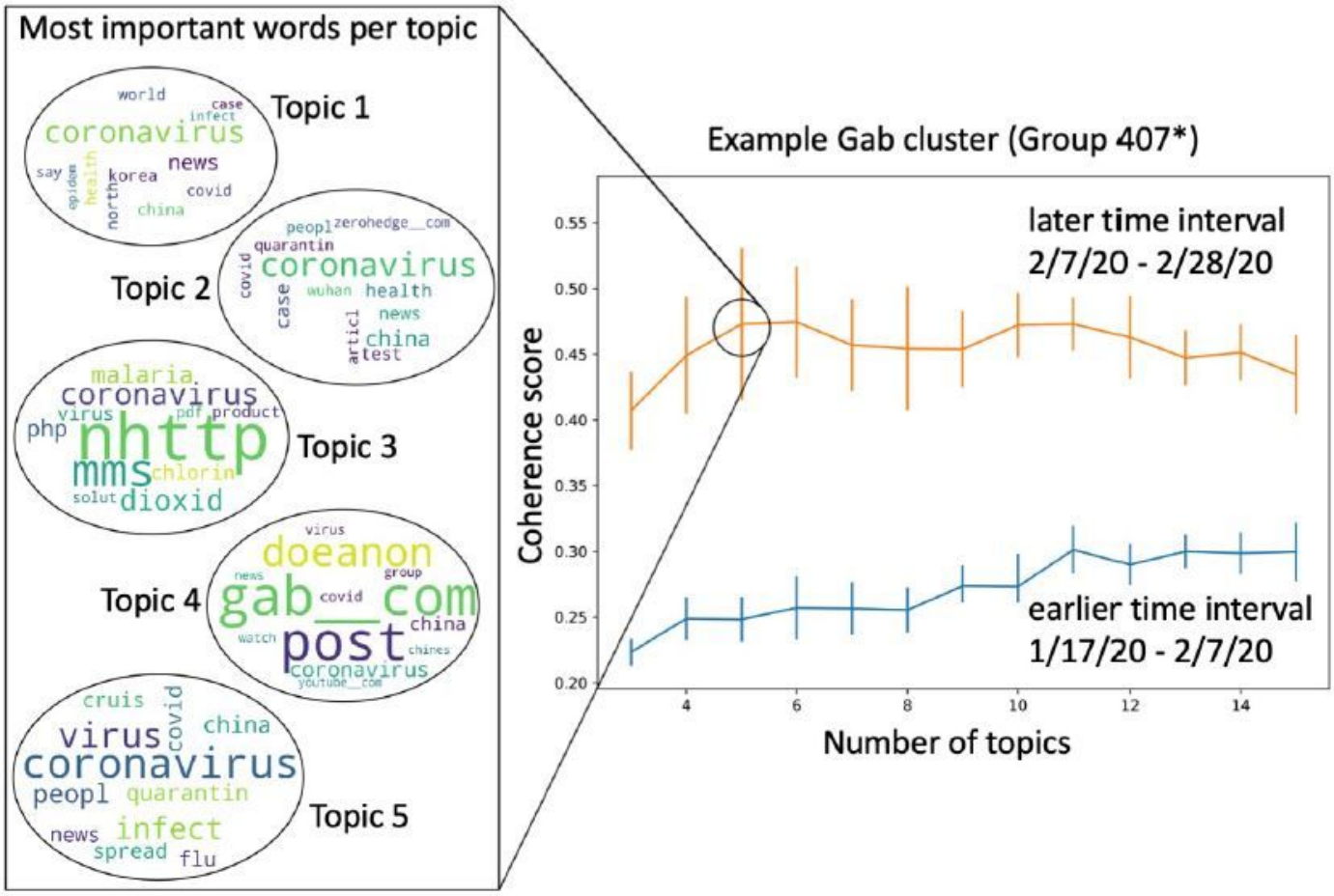

Figure 3

Evolution of COVID-19 content. Focusing on a single Gab hate cluster, this provides example output from our machine learning topic analysis of content. Even though discussion of COVID-19 only arose in December 2019, it quickly evolved from featuring a large number of topics with a relatively low average coherence score, to featuring a small number of topics with high average coherence score more focused around COVID-19. As the right-hand panel shows, the discussion of in this cluster became much more coherent, and focused on COVID-19, during the second three-week period we analyzed. The right-hand panel shows the keywords in each of 5 topics discussed on this cluster during that second three-week period. In the first three-week period, topics discussed featured profanity and hate speech such as f*** and n*****, but the conversation quickly become more focused and less like a stereotypical hate-speech rant. SI shows explicit examples of this content.

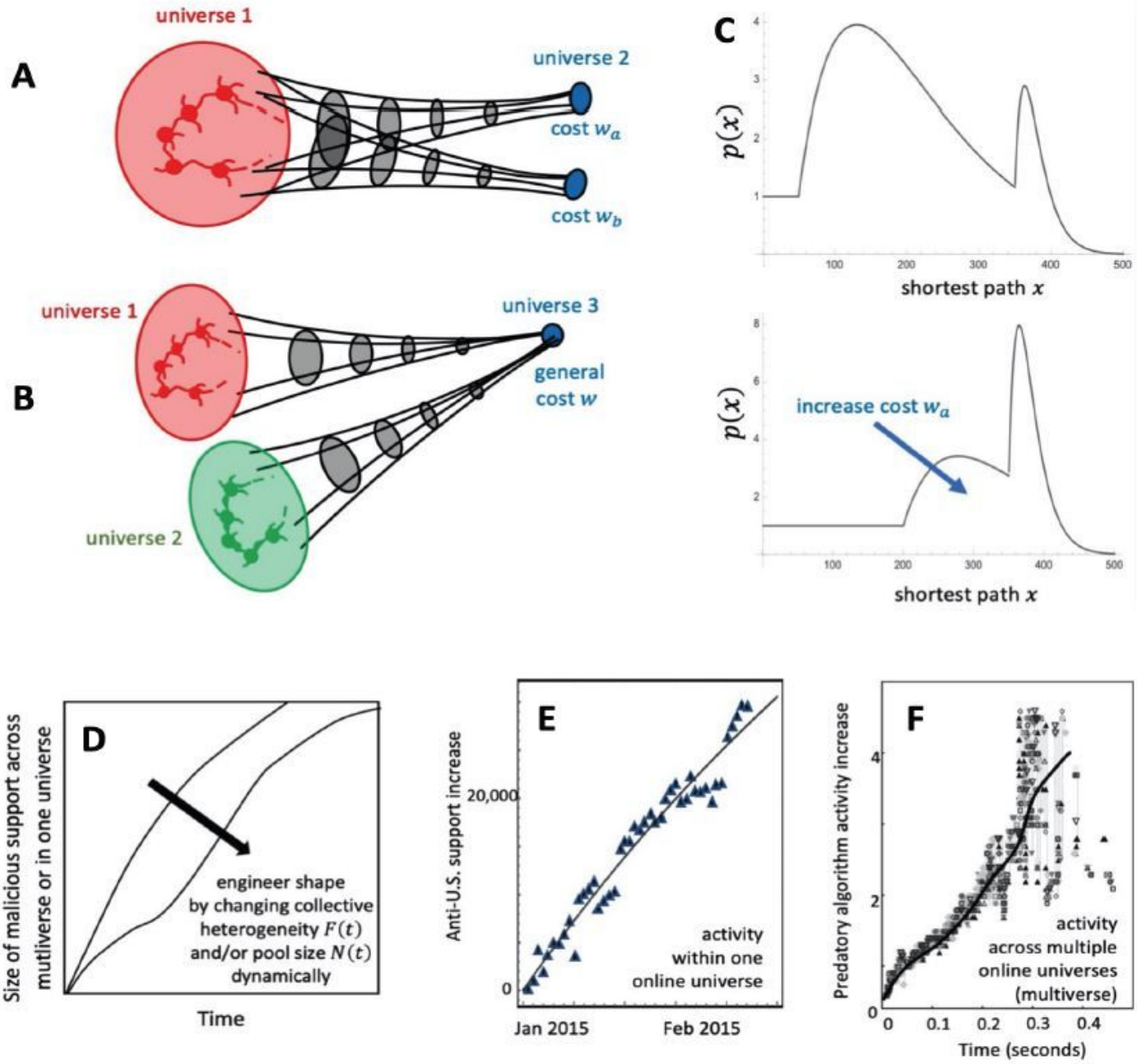

Figure 4

Wormhole Engineering to Mitigate Spreading. A-B: Typical motifs within the full multiverse in Fig. 2B. C: Mathematical prediction for motif A, showing that the distribution of shortest paths (top panel, shown un-normalized) for transporting malicious matter across a platform (i.e., universe 1) can be shifted to larger values (bottom panel) which will then delay spreading and will increase the chance that the malicious matter is detected and removed [25,26]. This is achieved by manipulating the risk that the malicious content gets detected when passing via the other platform: this risk represents a cost for the hate community in universe 1 when using the blue node(s). The same mathematics applies irrespective of whether each blue node is a single cluster or an entire platform, and applies when both blue clusters are in the same platform or are in different platforms. See SI for case B. D-F: Mathematical prediction that the

total online support for malicious matter can be manipulated by varying the online pool size of potential supporters N(t) and/or their heterogeneity F(t). The mathematics we develop here has implications beyond the hate network shown in Fig. 2B. E: Example of how an empirical outbreak of anti-U.S. hate across a single platform (VKontakte) in 2015 produces similar shape to upper curve in D. F: Empirical outbreak for the proxy system of predatory 'buy' algorithms across multiple electronic platforms [27] also produces a similar shape to lower curve in D. (See SI for details).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI11302020.pdf