IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 12, 2021, accepted December 26, 2021, date of publication December 27, 2021,
date of current version January 7, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3138982

How Social Media Machinery Pulled Mainstream Parenting Communities Closer to Extremes and Their Misinformation During Covid-19

NICHOLAS J. RESTREPO[1,2], LUCIA ILLARI[1,3], RHYS LEAHY[1], RICHARD F. SEAR[1], YONATAN LUPU[1,4], AND NEIL F. JOHNSON[1,3]

[1]The Dynamic Online Networks Lab, George Washington University, Washington, DC 20052, USA
[2]ClustrX LLC, Washington, DC 20007, USA
[3]Department of Physics, George Washington University, Washington, DC 20052, USA
[4]Department of Political Science, George Washington University, Washington, DC 20052, USA

Corresponding author: Neil F. Johnson (neiljohnson@gwu.edu)

This work was supported by the U.S. Air Force Office of Scientific Research under Award FA9550-20-1-0382 and Award FA9550-20-1-0383.

ABSTRACT We reveal hidden social media machinery that has allowed misinformation to thrive among mainstream users, but which is missing from current policy discussions. Specifically, we show how mainstream parenting communities on Facebook have been subject to a powerful, two-pronged misinformation machinery during the pandemic, that has pulled them closer to extreme communities and their misinformation. The first prong involves a strengthening of the bond between mainstream parenting communities and pre-Covid conspiracy theory communities that promote misinformation about climate change, fluoride, chemtrails and 5G. Alternative health communities have acted as the critical conduits. The second prong features an adjacent core of tightly bonded, yet largely under-the-radar, anti-vaccination communities that continually supplied Covid-19 and vaccine misinformation to the mainstream parenting communities. Our findings show why Facebook's own efforts to post reliable information about vaccines and Covid-19 have not been efficient; why targeting the largest communities does not work; and how this machinery could generate new pieces of misinformation perpetually. We provide a simple yet exactly solvable mathematical theory for the system's dynamics. It predicts a new strategy for controlling mainstream community tipping points. Our conclusions should be applicable to any social media platform with in-built community features, and open up a new engineering approach to addressing online misinformation and other harms at scale.

INDEX TERMS COVID-19, dynamical systems, misinformation, online, social computing, social media.

I. INTRODUCTION

Numerous studies have shown that social media helps feed the spread of misinformation and other harms [1]–[5]. Even before Covid-19, there were significant amounts of misinformation circulating every day—for example, against the measles vaccine [6], [7]. The pandemic further amplified this [8]–[10] because of the uncertainties surrounding Covid-19, and because people began interacting more online due to social distancing and remote working. Indeed, there was a huge jump in social media users during 2020 (13.2%) taking the total to 53.6% of the global population [11]. The top reason given by users for going online was to get information [11]. Online misinformation about Covid-19 has led to people losing their lives after rejecting vaccines and masks and drinking bleach. Furthermore, online misinformation about climate change is also now surging, with dangerous potential consequences.

Despite investing significant resources in policing their platforms, and suspending what they think are key accounts, social media companies such as Facebook still struggle with a daily deluge of new material to monitor [12]. At the same time, there are increasingly impatient calls from policymakers and governments for social media platforms to do 'more'.

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

But without some system-level understanding of the online misinformation machinery at scale, what does 'more' actually mean?

Here we address this urgent need for a mechanistic, dynamical understanding of how misinformation thrives across the social media system—just as would be demanded when troubleshooting problems in any other large-scale, hybrid engineering system that mixes hardware, software and humans. The complexity lies in the fact that the machinery of social media is built around encouraging people to connect into communities around some shared interest (e.g. Facebook pages built around parenting topics including family health) and then having these communities connect to each other and so on. Hence it is an intertwined system that blends the business model of the social media platform with that platform's community-building and community-connecting features, and the collective behavior of humans at scale.

We focus on Facebook since it is the largest and most widely used social media platform. Our main unit of analysis is the in-built community which is a social media platform feature (e.g. a page on Facebook) that allows people to group together online with the purpose of discussing some particular shared interest, e.g. parenting. Such in-built communities are a key driver of collective social activity on most social media platforms, including Facebook. Recent studies have shown that people (e.g. parents) increasingly rely on such communities as a source of information and advice concerning their families' health [13]–[15]. They build up a level of trust in the community to which they belong, and hence will be more likely to pay attention to its collective advice and information since it comes from peers having the same interests, e.g. other parents who are similarly worried about particular health issues or choices for their young child and who are willing to share their own personal experiences and opinions [13]–[15].

The consequence is that misinformation that happens to be circulating within and between these in-built communities, can influence not only these parents' decisions about their own daily practices and behaviors, but also their knock-on decisions about their children and their advice to other family members such as older parents. This includes the choice of whether to wear masks and have vaccines and booster shots—and in the case of opinions about the ability to control climate change, whether to make a determined effort to recycle goods and purchase energy-saving devices. Hence there is a large potential amplifying factor in terms of real-world impact from misinformation in these online communities, that stretches to others who may not even have an online presence themselves.

Our focus on online communities also brings an advantage of scale. Since each online community (e.g. Facebook page) can contain up to a million users or more with an average size of around 100,000, then our study of the online ecosystem of approximately 1000 communities offers insight into the collective behavior of roughly $100000\,(1000) = 100$ million users. Hence our study provides a unique view of the system at scale that goes well beyond existing case studies focused on small sets of online actors.

These online communities produce and share (emit) content and also receive content from other communities to whom they are connected. If a community A links to community B (e.g. Facebook page A 'likes' Facebook page B), this creates an information conduit from B into A and hence can expose A's users to B's content (e.g. posts from Facebook page B appear on Facebook page A). Hence understanding communities and their connections at the system level is crucial. An additional advantage of our study's focus on communities rather than individuals, is that it avoids any issues concerning access to personal information.

Given the above, this paper focuses primarily on mainstream parenting communities and how they are connected into more extreme communities from which misinformation originates, such as those built around anti-vaccination as well as more traditional conspiracy theories surrounding climate change, fluoride, chemtrails and 5G. By ''mainstream parenting communities'' we mean these are not communities that are actively promoting any anti-vaccination or other conspiracy views but are instead focused on everyday issues that likely preoccupy most parents with the means and ability to be active on Facebook, such as how many hours of screen-time and television should they allow, how to get their children to eat more vegetables, and what educational choices might be best. Their concerns also stretch to health of course—hence many have become observers of the online health debate around vaccines (i.e. neutrals as discussed in Sec. II). Also, we will use the term ''guidance'' in this paper interchangeably with information and advice, since guidance is defined in the Oxford Dictionary as ''advice or information aimed at resolving a problem or difficulty''.

The open question is then: What it is about how these mainstream communities are interconnected, and their potential interaction with other more extreme communities online, that makes the engineering problem of tackling widespread misinformation so hard to solve? And what might be done to overcome it at scale?

Our results in Secs. II-V start by unraveling the granular features of the relevant Facebook community machinery at scale, and how it evolved during the Covid-19 pandemic. These findings go well beyond our previous study in Ref. 7 which ended before Covid-19 and contained no such granular structure or dynamical details. The engineering-inspired analysis that we develop, shows mainstream parenting communities getting pulled closer to extremes, and the misinformation that they produce, during Covid-19. It reveals a strengthening of the bond between mainstream parenting communities and pre-Covid conspiracy theory communities that promote misinformation about climate change, fluoride, chemtrails and 5G—and it shows that alternative health communities acted as the critical conduits. An adjacent core of tightly bonded, anti-vaccination communities injected additional Covid-19 and vaccine-specific misinformation. We explain why this resulting two-pronged machinery can generate new pieces of

misinformation without needing any new news. We also show why Facebook's own scheme to supply reliable information about vaccines and Covid-19 was not efficient, and why targeting the largest communities does not work. We provide a simple yet exactly solvable mathematical theory for the system's dynamics, that predicts a new strategy for controlling mainstream community tipping points. It is inspired by existing analyses of dynamical systems from across the engineering sciences.

The data collection and network construction are discussed in Sec. II, while the main empirical results and analysis are in Sec. III. Section IV derives a mathematical theory of the system dynamics. Section V contains limitations of the study and Sec. VI summarizes the main conclusions.

The main contributions of this paper are as follows.

- We show that mainstream parenting communities on Facebook were subject to a powerful, two-pronged misinformation machinery during the pandemic, that pulled them closer to extreme communities and their misinformation.
- Our mapping of the online ecosystem shows why Facebook's own effort to post reliable information about vaccines and Covid-19 was not efficient, why targeting the largest communities will not work, and how this machinery can generate new pieces of misinformation perpetually.
- We provide a simple yet exactly solvable mathematical theory for the system's dynamics. It predicts a new strategy for controlling mainstream community tipping points and should be applicable to any social media platform with in-built community features.
- Our results open a new engineering approach to addressing online misinformation and harms at scale.

II. DATA

Figure 1 shows a flowchart of our data compilation process. It extends the approach presented in Ref. 7. We start by using keyword searches to collect a list of Facebook communities (pages) surrounding the health debate over vaccines on Facebook [7]. This produces an initial core set of Facebook communities (pages). Then we see which pages they connect to (i.e. follow) and add these to the list. We then remove any pages that are not a self-organizing community of users, e.g. we remove businesses. Hence we obtain a final list of communities, each of which is represented as a node in our subsequent network analysis, and the links between them. These steps can be carried out automatically, but the lists obtained are in any case checked manually for errors. This online vaccine debate broadened post-Covid to also include the topic of Covid-19 and vaccines during 2020. Over the period of study from the end of 2019 (i.e. pre-Covid) to the end of 2020, the number of communities changed only slightly. The overall number is 1356 nodes (communities) with 7154 links between the nodes in the largest network component at the end of 2020. We checked that our main conclusions are robust to errors in this data collection process, by randomizing and also removing up to 10% of the nodes and links and repeating our analysis.

The subject-matter experts in our team then classified this final list of nodes (communities) as 'pro' (i.e. pro-vaccination), 'anti' (i.e. anti-vaccination), or 'neutral' (i.e. they had not shown a specific preference). The pro and anti classifications require that either (a) at least 2 of the most recent 25 posts dealt with the pro-vaccination or anti-vaccination debate, or (b) the page's title or ''about'' section described it as pro-vaccination or anti-vaccination. Then they further sub-categorized the neutral communities into types (e.g. parenting). The subject matter experts had each had several years experience in analyzing and classifying online community content on Facebook and other platforms. At least two researchers classified each node (community) independently. If they disagreed on their suggested classification, a third researcher reviewed the content and then all three reviewers discussed these cases. Agreement was reached in each case. The self-weeding tendency within Facebook pages tends to reduce content from bots and also fake profiles. We kept the present study focused on English, though this can be easily generalized using our same procedure. Beyond that, our study was global and not limited to a particular region.

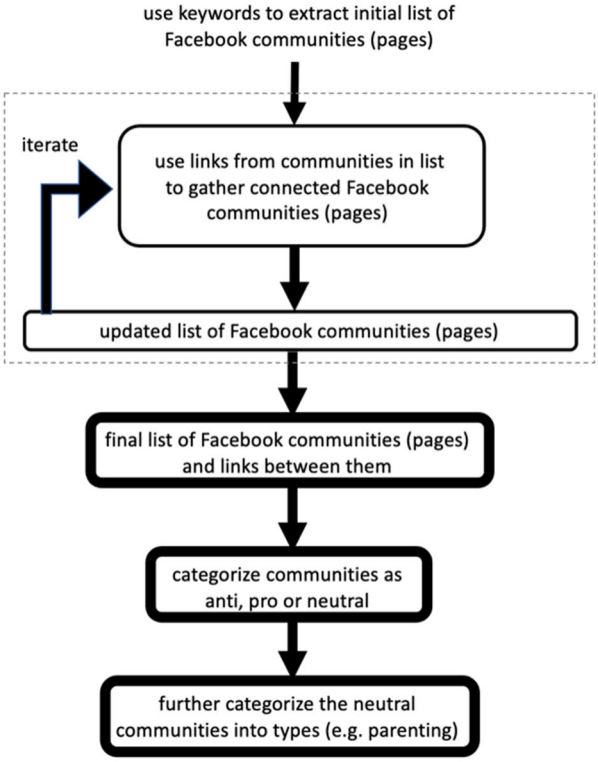


FIGURE 1. Flowchart of our data compilation process.

There are other possible approaches to collecting such a list of Facebook communities and their interconnections, and hence nodes and links for subsequent network analysis.

However, many of these other approaches have significant drawbacks in our opinion:

(1) Some studies rely on lists of communities obtained using CrowdTangle, which is a commercial application tool owned by Facebook. But researchers outside Facebook have little knowledge or quantitative explanation of how and why this tool returns the results that it does. In other words, it is effectively a black-box tool, which makes it unacceptable for academic science research in our opinion. Researchers outside Facebook do not know how the results that it returns depend on Facebook's secret algorithms, architecture and databases, or how all this changes in time. Nor can researchers outside Facebook control any of this: they simply input search prompts and the tool spits back data. Nor do they know about the completeness of the black-box search results returned. Nor do they know if there is any bias in the search process within the black-box tool, nor what that bias might be or how big it is. Our own investigations suggest that similar searches can produce quite different results. While a larger list may indeed be obtained using such a black-box tool, that list may be significantly biased and hence less reliable than a smaller sample obtained using a non-black-box tool. The CrowdTangle search output is also not very precise. It can for example include results with different spellings that are unrelated in topic. Nor are the numbers that it returns checkable or proven to be accurate, which further calls into question the reliability of studies that use it for quantitative academic analysis. Nuancing the search terms can produce very different results, adding to concerns about how complete and robust the output is for academic research. Also, doing searches about the past cannot easily reveal communities that have removed themselves or were removed by Facebook, or have changed their name. It therefore remains unproven that such black-box tools are suitable for rigorous, reproducible scientific research as opposed to simply being used as a search tool for businesses and for qualitative exploration of a particular story. Without systematic, quantitative studies against ground truth lists, one cannot assume that findings obtained using black-box tools are reliable. Nor is the number of candidate communities that emerges an indicator of a larger sample and hence a broader or more reliable study, since it is always possible to capture many more communities by using a coarser net to capture many less relevant and potentially biased examples.

(2) After classification of the communities, whether through CrowdTangle or otherwise, different studies with slightly different classification schemes may end up with very different numbers of communities in a given category, e.g. more anti communities. Again, this does not mean that a study with a larger list is better or has more reliable results, since the classification criteria are not identical for the categories, i.e. starting from the same bag of candidate communities, the researcher-chosen criteria for the 'anti' label in any given study could simply allow for more objects from that bag to be assigned the anti label.

(3) Other studies may use different ways of defining links between nodes (communities), e.g. URLs listed in the content. But there is no guarantee that these URLs represent any meaningful connection between the majority of users in one community and another, nor that it influences their subsequent behavior in any significant way. This needs to be proven before any subsequent network analysis can be regarded as meaningful. By contrast, in our study the links between nodes (communities) are better defined, i.e. if page A 'likes' page B then this community A links to community B which creates an information conduit from B into A and hence exposes A's users to B's content (e.g. new posts from Facebook page B can appear on Facebook page A).

(4) Though we only focus here on Facebook pages, Facebook groups also exist as a separate in-built feature on Facebook. However, Facebook groups tend to be relied on for more private conversations, and private Facebook groups cannot be openly accessed. So, studies that include public Facebook groups can be problematic, since most of the interesting content is private. We find that a significant fraction of the Facebook pages that we study act as a public-facing vehicle for such private Facebook groups, in that a link to the private group appears within the public page (see Supplementary Information (SI) for an example). Hence the network of Facebook pages that we study likely acts as the crude skeletal structure around which the full but largely hidden network of public and private Facebook groups actually operates. Therefore, our study of the network of Facebook pages (communities) does indeed give insight into the shape of the full online ecosystem on Facebook.

In our study, the building of the list in Fig. 1 can be facilitated by automated data collection using scripts in R or Python, but can also be done manually since the final number of nodes is relatively small. The collection of the links between them can also be done manually, though obviously it is speeded up if a script is used. But we avoid using a third-party tool such as CrowdTangle because of the reasons given above, i.e. the difficulty in knowing exactly how it produces the results that it does and hence its reliability for scientific investigation. The publicly available software Gephi was used to plot the network diagrams associated with Figs. 2-5. The software Mathematica was used to plot the solutions of the differential equations in Fig. 6. We stress that our study goes well beyond the analysis of Ref. 7 because it examines the impact of adding the topic of Covid-19, by including data through 2020; and it examines the granular identity of the neutral communities' interests, and hence classifies them beyond the single 'neutral' label.

III. ANALYSIS AND RESULTS

Our categorization process summarized in Sec. II and Fig. 1 yields a network of interconnected communities (nodes) with 211 'pro' communities (blue nodes in Fig. 2) comprising 13.0 million individuals, whose content actively promotes establishment health guidance (pro-vaccination); 501 'anti' communities (red nodes in Fig. 2) comprising 7.5 million individuals, whose content actively opposes this guidance (anti-vaccination); and 644 'neutral' communities (green

nodes in Fig. 2) comprising 66.2 million individuals, that had community-level links with pro/anti communities pre-Covid but whose content is focused on other topics such as parenting, pets, organic food, and who have not expressed a stance. There are 7154 links between communities (nodes) in the largest network component shown in Fig. 2.

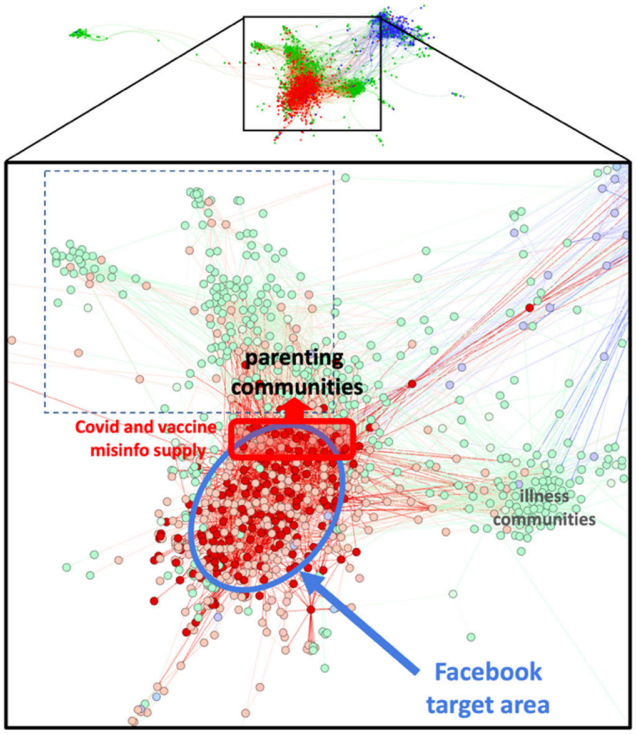


FIGURE 2. Ecosystem of Facebook communities in December 2020. Each node is a community (Facebook page). Red (blue) nodes are communities that were anti (pro) establishment guidance about vaccines before Covid-19: they then broadened their content to include the topic of Covid-19 during 2020. The 644 neutral communities (green nodes) totaling 66.2 million users, are not focused on such topics but are entangled with other communities that are. Illness communities focus on supporting sufferers of long term, non-Covid illnesses, e.g. autism. The ForceAtlas2 layout that we use throughout this paper, means that the proximity of nodes reflects more links and hence a higher chance of those nodes sharing the same content and hence the same misinformation. Darker shaded nodes are communities that displayed the Facebook banner promoting establishment guidance: nearly all of these are antis (red) and lie within the blue ring. Lighter shaded nodes are communities that did not display the Facebook banner.

We checked that our main conclusions are robust to errors in the categorization process, by randomizing or removing up to 10% of the categories. We also note that although our data collection in 2019 was focused around the vaccine debate, the communities that we found talking about vaccines prior to Covid-19 then broadened their discussions to include vaccines and Covid-19 afterwards. In particular, the anti communities expanded their narrative from promoting misinformation about vaccines prior to Covid-19, to promoting misinformation about Covid-19 and vaccines afterwards.

In the analysis in this paper, we plot the networks of these interconnected communities (i.e. nodes, each of which is a Facebook page) using Gephi's ForceAtlas2 layout. This ForceAtlas2 layout algorithm follows physical rules by treating the nodes as balls and the links as springs, and then letting the system relax by itself. This means that the visual appearance of the resulting network layout is completely spontaneous. It also makes the resulting layout easy to interpret since sets of communities (nodes) that are more interconnected will appear closer together, and hence will be the ones more likely to have shared content including misinformation. The Supplementary Information available online (SI) demonstrates this explicitly, by showing quantitatively how the resulting network's spatial appearance follows directly from its links in ForceAtlas2, and how quantitative information about the aggregated strength of links can be inferred from the layout, albeit in an approximate way.

Figure 2 (top) shows the full network, which is an end-of-2020 version of the pre-Covid 2019 one that was presented in Ref. 7. The panel below then zooms in on the portion of this network that contains the highly interconnected anti communities (red nodes). Given that the ForceAtlas2 network layout is spontaneous and agnostic to node type, the segregation that emerges between the anti communities (red nodes) and the pro communities (blue nodes) is striking, as is the high level of anti-neutral (red-green node) entanglement. It leads to the neutrals being concentrated near the antis, hence the magnified portion includes nearly all the neutrals. By contrast, most of the pros are heavily connected to each other, which means they are primarily sharing guidance with each other and hence effectively 'preaching to the converted'. Though this tendency already existed in 2019, it is surprising that the pros did not manage to make any discernible improvements to their segregation during 2020.

Clearly this segregation between antis and pros, together with the close proximity between the large set of neutrals and the antis, must be taken in to account by any public health or policymaking initiative that aims to expose the population to establishment guidance, including best science advice and information about health. Also, we note that when analyzing the URLs within the community content, we find that different types of nodes tend to link to outside news sources of different types and in different amounts. This in turn serves to refresh their content continually and suggests that any fact-checker or 'inoculation' approach to tackling misinformation among the antis and neutrals, will be very hard to maintain and update in real time and at scale.

In addition to the clustering together of antis and separately of pros, Fig. 2 also shows there is clustering together of neutral communities of a given type (e.g. mainstream parenting communities). Since the ForceAtlas2 network layout is agnostic to node type, this clustering serves as a demonstration that the network links that we define and identify in our study are indeed meaningful. Specifically, particular types of neutrals cluster together because they are more highly interlinked. For example, mainstream parenting communities are highly interlinked with other mainstream parenting communities. This suggests that parents prefer seeking and sharing advice among themselves (i.e. other parents) and do

so not only within their own parenting community, but also across parenting communities [13]–[15]. The same holds for the conspiracy communities that focus on non-Covid and non-vaccine topics (e.g. climate change, fluoride, chemtrails, 5G) but which are neutrals since they do not display an anti-vaccine stance. Illness communities—such as communities of people whose lives are affected by cancer, Parkinson's disease or autism, but which classify as neutral—also end up clustered together.

From a broader engineering perspective, the clustering behavior in Fig. 2 is an explicit example of collective, self-organized behavior emerging at new scales within a complex dynamical system. The underlying system comprises interacting, heterogeneous objects (humans) at the smallest scale, which then self-organize into communities at a higher scale. Then treating each community as a renormalized object (node) at this higher scale, these higher-level objects (nodes) interact with each other to form self-organized clusters-of-communities (clusters of nodes) at an even higher scale. As we show later, we can then take this to an even higher scale by treating such clusters-of-communities as renormalized nodes (super-nodes) that then interact with each other. This renormalization approach will allow us to develop a tractable mathematical description of the system dynamics, as shown in Sec. IV.

The darker nodes of a given color in Fig. 2 show the communities that received one of Facebook's own promotions against misinformation, i.e. they received a banner on their Facebook page promoting official information sources (e.g. Centers for Disease Control (CDC)). The SI shows an explicit example of this. Nearly all of the targeted nodes are antis (red nodes). Among all the anti communities, only 39% received the Facebook intervention. Moreover, most of these are confined in a small core region (darker red nodes inside the blue ellipse). This left most other communities, including all the mainstream communities (green nodes) focused on parenting, without this health guidance. Overall, less than 2% of neutral communities received the Facebook banner.

This is concerning for two reasons: (1) As emphasized in Sec. I, parents tend to turn to such Facebook communities for guidance on issues such as their family's well-being, and did so particularly during the pandemic. (2) It has recently been shown [16] experimentally and theoretically that an online community can suddenly tip (e.g. adopt a certain piece of misinformation as true) in a reproducible way if there is a committed minority of around 25%. Since all the nodes are interconnected in Fig. 2 and most nodes did not receive Facebook's banner promoting establishment guidance, this enhances the risk of such tipping events cascading quickly across the ecosystem. The dotted square region that we highlight in Fig. 2, is of particular concern since it includes ≈30 million users and contains mainstream parenting communities as well as those promoting long-standing non-Covid and non-vaccine conspiracy theories, primarily around climate change, fluoride, chemtrails and 5G.

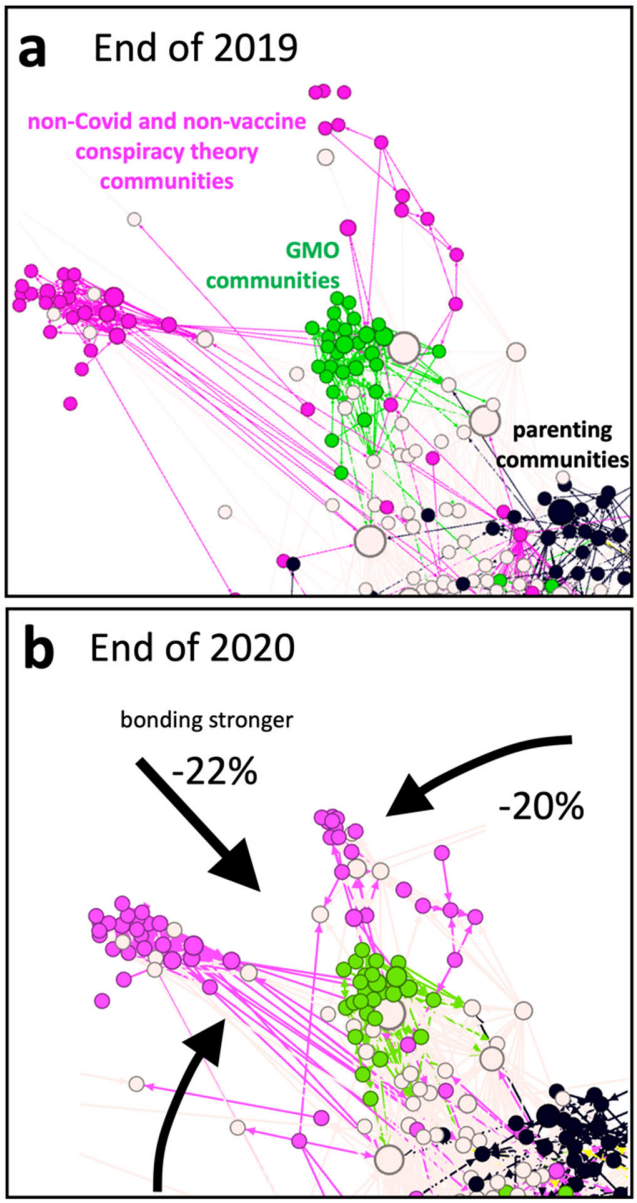


FIGURE 3. Parent–conspiracy–theory bonding strengthens during Covid. Dotted portion from Fig. 2 shown (a) just before Covid, and (b) just before the rollout of Covid-19 vaccines, using same scale in both (a) and (b) to enable comparison. Neutral nodes, which are all green in Fig. 2, are shown here as purple, light green and black to denote their different types (e.g. black are mainstream parenting communities). The (light) red nodes are antis. Between the end of 2019 (panel a) and the end of 2020 (panel b), the distance between non-Covid/non-vaccine conspiracy theory communities and parenting communities shortens by approximately 22% and the angle reduces by approximately 20%.

Figure 3 examines in detail this dotted square region from Fig. 2, with different types of neutrals shown as different colors. As in a real molecule, the forces in the ForceAtlas2 layout pull together clusters of nodes (i.e. clusters of communities) when the overall linkages between them increase, i.e. the bond is strengthened and shortens. We therefore borrow from chemistry by defining a bond length as the

distance between the centers of two clusters of nodes, and the bond angle as the angle between two bond directions. During 2020, the bond lengths between the two clusters of non-Covid/non-vaccine conspiracy theory communities and the parenting communities shortened by approximately 22% on average while the bond angle decreased approximately 20% from 39 to 31 degrees (see SI). We recognize that these numbers are just crude proxies for the bonding, and that just as in chemistry it is hard to measure precisely the midpoint positions from which to determine angles and bond lengths. Hence these numbers are just rough estimates. We also know that one should not read too much into network layouts. But as we show explicitly in the SI, the relative changes in network layouts within the ForceAtlas2 algorithm can indeed be quantified and interpreted in terms of changes in the total link (and hence bond) strength and length.

We have therefore identified the first prong of the misinformation machinery impacting mainstream parenting communities during the pandemic, i.e. the strengthening and hence shortening of the bond between mainstream parenting communities and pre-Covid conspiracy theory communities that promote misinformation about climate change, fluoride, chemtrails and 5G. This is a significant finding for understanding the spread of misinformation since the ForceAtlas2 layout means that sets of nodes (communities) that are visually closer to each other in the network will tend to have more links between them, and hence will have a higher chance of sharing the same content and hence sharing the same misinformation.

Interestingly, the origin of this bond strengthening mechanism does not lie in an obvious place, i.e. it does not come from the numbers of direct links from mainstream parenting communities to non-Covid/non-vaccine conspiracy theory communities, nor to communities against genetically modified foods (GMO). Indeed, there are zero such direct links. Instead, it is the alternative health communities that act as the interconnecting bridge and hence conduit between them (Fig. 4a). These alternative health communities promote, discuss, and/or feature content about alternative cures and practices, including homeopathy, naturopathy, and spiritual healing, as opposed to modern medical practice. The third-party bonding that they mediate, mimics 'superexchange' bonding in complex biochemical molecules [17]. We recognize that applying a simple network measure such as betweenness centrality to such a multi-partite network will not properly quantify its complexity, nevertheless its popularity as a network measure makes it worth analyzing. Betweenness centrality is a standard measure of a node's capacity to act as a conduit. We find that as a result of the rewiring during 2020, these alternative health communities showed up to a 40% increase in their betweenness centrality.

The posts in these alternative health communities are not generally about conspiracy theories or vaccines, confirming that this is not their overall focus or intention—but deep in the content of the replies to the comments on the posts, one can see new conspiracy theories and misinformation being continually generated by blending themes from broader conspiracy theories, e.g. text in Fig. 4a combines narratives about World War III, 5G, vaccines, oil rigs, and vitamins.

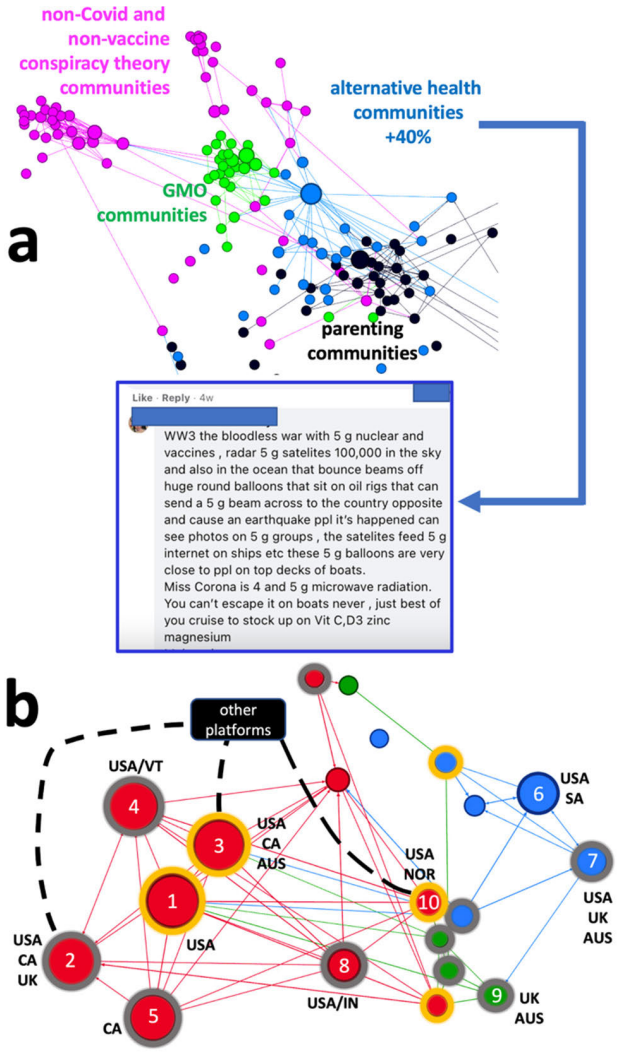


FIGURE 4. Panels a and b show the 2 prongs of the misinformation machinery that impacted mainstream parenting communities during Covid-19. a: First prong. Alternative health communities, which focus on the power of the immune system, provide the key bonding mechanism during 2020 between mainstream parenting communities and non-Covid/non-vaccine conspiracy theory communities. Each node shown is one of the green nodes in Fig. 2. b: Second prong. It comprises a highly interconnected and active core of under-the-radar anti communities which lie in the red box from Fig. 2, just below the mainstream parenting communities. b shows the subnetwork from Fig. 2 that comprises the top 20 communities from among all those in Fig. 2, ranked by betweenness centrality, i.e. ranked by ability to act as a conduit for (mis)information and conspiracy theories. SI lists these nodes' properties. Their ranking is shown for the top 10 nodes. The top 5 are all antis and are highly interconnected, and hence form a powerful misinformation 'sub-engine' sitting just below the mainstream parenting communities (red box, Fig. 2) that supplies them with Covid and vaccine misinformation. The countries where the administrators of each node (Facebook page) are located, are also shown. A yellow ring denotes a node that represents a potential net emitter of misinformation, i.e. a node for which the number of other nodes that link to it, and hence which can receive content from it, minus the number of other nodes that it links to, and hence it can receive content from, is positive. A gray ring denotes a potential net receiver of misinformation, which is when this overall number is negative (see SI).

Overall, they seem to make such narratives more palatable to mainstream audiences by appealing to the basic instinct of protecting one's family against perceived future threats in any way possible. Psychologists have long known that people are susceptible to taking on board misinformation [18], including online misinformation related to health and policy [19]–[21].

The second prong of the misinformation machinery impacting mainstream parenting communities, emerges from Fig. 4b. Figure 4b shows the 20 nodes with the highest betweenness centrality within the entire network in Fig. 2 (see SI for explicit numbers). It also includes any links between them from Fig. 2. Though they have the highest betweenness centrality within the entire network, it is surprising that these top 20 nodes are linked with each other as much as they are, and what color (type) of nodes they are. Specifically, Fig. 4b shows that the top ranked nodes are all antis and that they are all highly linked to each other. The numbers on the nodes show their ranking by betweenness centrality among all the 1356 nodes in the full network comprising pros (blue nodes), antis (red nodes) and neutrals (green nodes) in the Facebook ecosystem. These highest-ranked antis hence form a 'sub-engine' that enables highly efficient sharing of misinformation across communities, since their high betweenness centrality means they can each act as highly efficient conduits to lower ranked nodes in Fig. 2, as well as being highly connected to each other and to other high betweenness pros and neutrals in Fig. 4b. Moreover, the antis in this sub-engine in Fig. 4b sit just below the mainstream parenting communities in Fig. 2. Given the fact that closer proximity in the ForceAtlas2 network layout favors more sharing of material and hence more sharing of misinformation, this sub-engine would have been able to continually and efficiently pump Covid and vaccine-specific misinformation into the mainstream parenting communities throughout 2020 (see SI for node details).

Remarkably, the robustness and resilience of this second prong of the misinformation machinery (Fig. 4b) does not originate from the size of the anti communities within it: i.e. the antis in Fig. 4b are not the largest antis despite being the best connected. Indeed, their relatively small size can explain how they managed to operate without being shut down by moderators during 2020, i.e. they were simply below the radar. By contrast, the largest anti communities are not the best connected (see SI for numbers) i.e. they have much lower betweenness centrality.

Specifically, the top-5 anti communities ranked by betweenness centrality (Fig. 4b) are only ranked 54th, 72nd, 64th, 473rd and 248th by size (see SI). Overall in the full network, not only do each of these 5 have more than 100 links, each of them has a huge difference between the number of inbound and outbound links: $+139, -107, +67, -121, -118$ yielding z-scores of 59, -48, 28, -54, -52 compared to a network with randomized links. Such a large positive (or negative) value means each of them has the potential to act as a strong net emitter (yellow ring in Fig. 4b) or net receiver (gray ring in Fig. 4b) of misinformation, and that they can fit together like lock and key akin to ions with opposite valences in chemistry. This adds to their under-the-radar resilience.

Furthermore, many of these nodes in Fig. 4b have administrators from across the globe, as shown. This helps give mainstream parenting communities, for example, the impression that any (mis)information being shared is endorsed globally and locally. It also helps them craft narratives so that they hold appeal across different continents and cultures, and yet also have local relevance. Furthermore, several have a simultaneous presence on other platforms, as indicated by the links in Fig. 4b. Indeed, their content shows them directing their users to regroup on other platforms such as MeWe, Parler, Gab and Telegram, with the purpose of continuing their conversations away from Facebook's moderators (see SI for an example).

There are three important broader takeaways from Fig. 4b. First, it calls into question any moderation approaches that focus on the largest and hence seemingly most 'visible' communities, as opposed to the smaller ones that are better embedded. Second, it warns against pinning the problem of online misinformation on a top-10 or top-12 list compiled according to 'visibility', despite this being a popular narrative in the media. Third, it suggests that the key to an adversarial (e.g. anti-establishment health guidance) network's long-term survival, does not lie in having several high powered individual nodes, but rather by it adapting to develop a self-organized 'strength in depth' where many relatively minor nodes (in terms of size) develop a high individual betweenness centrality and hence ability to act as a conduit for (mis)information, and in addition they also become highly interconnected between themselves. They hence sit under-the-radar in terms of size, but get to efficiently promote misinformation by being so well interconnected. This is a complex network generalization of the popular saying that it is not the strongest that survive, but the most adaptable. It also hints at a better way to think about and hence tackle other online harms such as hate.

So what might happen if Facebook were to cut this machinery off from outside information sources and restrict its user base? One possibility is that the remaining users would then use freely available, off-the-shelf text-generating algorithms such as GPT-2 to autonomously generate high volume streams of text narratives that look like they were entirely human made [22]. To demonstrate the serious nature of this threat concerning online misinformation in the current context, we provide examples of such AI-generated texts of Covid and anti-vaccination (mis)information in the SI which we generated using GPT-2 by inputting prior online narratives from the anti communities. Even to the expert eye, these AI-generated texts have a fresh, human-like appearance. This suggests that the remaining anti community users can indeed continue to produce and circulate fresh (mis)information indefinitely even if cut off entirely from outside information sources, i.e. GPT-2 and its more powerful successors can make up for being cut off from outside information sources and from restrictions on their user base and hence posting-power. Furthermore, the rapid way in which large volumes of

fresh, human-like texts can be automatically generated, can be used to stay one step ahead of moderation schemes that detect the spreading of already known pieces of misinformation. Similarly, they can also remain one step ahead of bot-detection schemes that rely for their performance on tell-tale machine signatures such as repetitive patterns in the output and online activity.

We have therefore identified a dangerous, two-pronged misinformation machinery that has developed during the pandemic. It brings non-Covid and non-vaccine conspiracy theory communities closer to mainstream parenting communities (prong 1) while simultaneously feeding them Covid and vaccination misinformation (prong 2). Clearly, combatting online conspiracy theories and misinformation cannot be achieved without considering these multi-community sources and conduits. Our findings also suggest why Facebook's promotion of information banners in Facebook pages (Fig. 2) failed to stop the mainstreaming of conspiracy theories and misinformation, because their targeting was limited to an inner core (Fig. 2) and because most of the interconnected mainstream communities and non-Covid/non-vaccine conspiracy theory communities lie outside this. Targeting the largest individual communities will also not work since the major conduits do not involve the largest nodes (Fig. 4b). Finally, we note that as a simple consistency check of our finding that alternative health communities acted as critical conduits, we performed a machine-learning analysis of 60 non-Facebook websites during August, September, and October 2020. Our sentiment analysis used the lexicon of positive and negative words provided by Liu [23] with net sentiment defined as the number of negative words subtracted from the number of positive words (see SI for results). The website with the highest level of positive sentiment was indeed alternative health. We also note that more generally, our findings are consistent in spirit with Ward *et al.*'s call for analysis of the granular details of the actual communities to which people belong [24], [25] and hence complement the many existing studies focused more toward individual behavior [26]–[36].

IV. TOWARD A DYNAMICAL SYSTEMS MODEL

Having understood these empirical aspects of the system, we now seek a mathematical model of the system dynamics—in particular, hidden system instabilities (tipping points). Our aim is to obtain a model that is the most transparent possible, yet which can also capture the overall trends in the data and hence add understanding about the mechanisms by which these patterns might have been produced. We do not seek absolute best numerical fits to the data: better fits can of course be achieved by including more parameters and using more sophisticated tools including machine learning. Yet such black-box machinery typically adds little additional mechanistic insight, and indeed may obscure the key underlying dynamical processes. Hence we will follow the simplest possible path, making approximations and assumptions that seem reasonable based on our empirical experience with the online system. We stress that all our steps, our approximations and our assumptions can be generalized at the expense of more complex equations and hence a loss in transparency. Our approach here is hence in line with existing dynamical modeling across physics and it yields equations akin to those seen across the engineering sciences.

The full system in Fig. 2 has too many objects and interactions to allow for any tractable mathematical description, hence we will simplify it as follows: (1) We aggregate nodes (i.e. communities) of the same type (e.g. all antis) into a super-node, following the renormalization approach to describing multiscale complex systems that we mentioned in Sec. III. (2) We interconnect these resulting super-nodes by weighted links whose weight is given by the total number of links between nodes of the relevant types. This yields a renormalized system containing a number n of supernodes interconnected by weighted links. The value of n can be chosen by the model-builder based on the desired trade-off between choosing a smaller n to decrease the complexity of the equations, and a larger n to increase granularity. For example, a model of interacting super-nodes comprising all antis, all pros and all neutrals aggregated together, would correspond to $n = 3$.

Figure 5 provides a visual overview of the tipping point analysis that follows, using this coarse-grained (i.e. super-node and weighted link) version of the full network from Fig. 2. Figure 5a shows the specific example in which the mainstream parenting communities' super-node is the focus. In addition, we choose to include a super-node for all antis (red box); a super-node for all pros (blue box); and a super-node for each of the 3 more prominent neutral sub-categories, while the rest of the neutrals are aggregated together into 'other communities'. The result of counting the links between the constituent communities then yields the weighted links between the super-nodes shown in Fig. 5a. The super-node formed from the 64 mainstream parenting communities (black box in Fig. 5a) has a weighted link (information conduit) of weight 13 from the alternative health community super-node and one of weight 80 due to the links among themselves, i.e. self-loop.

We now focus on obtaining a dynamical equation for the super-node comprising the mainstream parenting communities (black box super-node in Fig. 5a) though we stress that the formulae that we show below can be applied to any category of super-node and can be generalized to any other aggregation choice and any n.

We introduce a variable $G(t)$ to crudely capture the collective activity of the 64 mainstream parenting communities' super-node (black box in Fig. 5a) in the online debate over establishment health guidance. Given that these mainstream parenting communities would likely show very small relevant activity in this health debate network if they had zero concern over establishment health guidance, we can justifiably interpret $G(t)$ as the collective level of concern of the mainstream parenting community members (parents) over establishment guidance. To measure $G(t)$ empirically, we could for example measure the total number of their members or their relevant

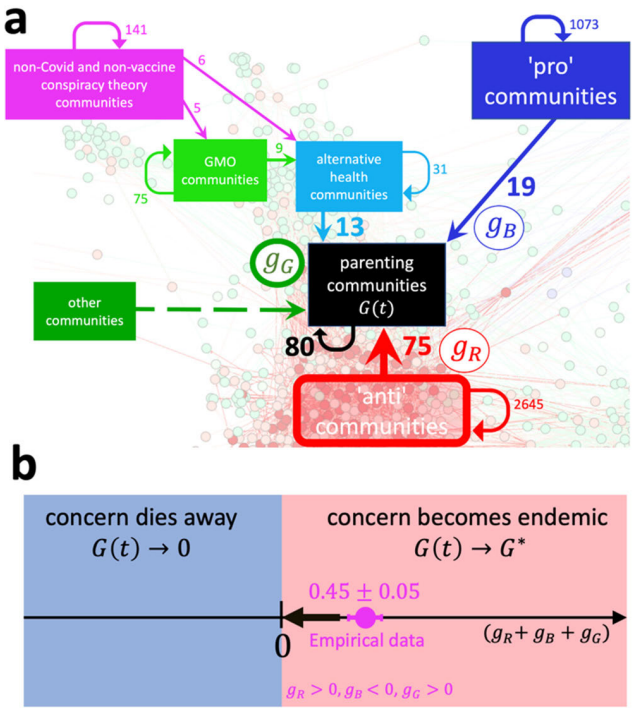


FIGURE 5. **Misinformation tipping point analysis using a coarse-grained version of the full network from Fig. 2. a: Weighted links (i.e. information conduits) between super-nodes of different community types, are obtained from the full network in Fig. 2 by aggregating all the nodes of a given type into a super-node, and summing all the links to get a weighted link between these super-nodes. Each link between communities in these super-nodes increases this weighted link by 1. Self-loops arise (e.g. 80 links between mainstream parenting communities, black box). $G(t)$ is an appropriate empirical measure of the collective activity of the super-node containing all the mainstream parenting communities (black box) at time t, chosen so that $G(t)$ crudely captures their collective level of concern at time t. b: Our model's mathematical prediction for future $G(t)$ using the sum of the couplings $(g_R + g_B + g_G)$. The numerical value shown is determined by fitting the model equations with $n = 3$ to the empirical data during 2020 (see Fig. 6) and is consistent with the value obtained independently from panel a (see text). Hence the prediction for the future based on current conditions, is that $G(t)$ (i.e. mainstream parenting community concern) will increase and become endemic, i.e. $G(t) \to G^*$. See text for a discussion of strategies to push the system over the tipping point (black arrow) so that $G(t) \to 0$ instead.**

postings across all 64 of their Facebook pages at time t. We leave the issue of the best empirical measure of $G(t)$ to another study, since we are focused here on developing the model and an analysis of its tipping point(s). Suffice to say it can always be measured, at least crudely, in one way or another.

In prior published work, we analyzed collective online behavior around other controversial topics (e.g. domestic extremism, jihadi extremism, far-right hate) in which online activity also grows and is measured through the number of members and/or number or toxicity of postings [37]–[39]. There we found that the growth in time of each community's activity, and also the communities as a whole, $x(t)$ (which played the analogous role to $G(t)$) followed a particular growth equation for a 'gel'. We also showed how this 'gel' equation can be derived from first principles mathematically by solving coupled differential equations for the aggregation process of online users. We refer to the supplementary material of Ref. 37 for the full mathematical derivation, which is not of concern here. What is important, is that the resulting gel growth equation is well approximated by the solution to a much simpler growth equation $\dot{x} = a(x_0 - x)$ where a is the growth rate and x_0 denotes a capacity. We will therefore adopt similar linear dynamical forms here for the n super-node equations.

Given this, we can write down a minimal model for the time-evolution of the mainstream parenting communities' super-node $G(t)$:

$$\frac{\partial G}{\partial t} = g_R\,(R - G) + g_B\,(B - G) + \sum_i g_{G_i'}\,(G_i' - G) \quad (1)$$

which captures the fact that $G(t)$ will in general depend on the corresponding levels of all other super-nodes, i.e. $G(t)$ depends on $R(t)$ for the anti communities' super-node, and on $B(t)$ for the pro communities' super-node, and on $G_i'(t)$ for the other neutral category super-nodes, where $i = 1, 2, ..$ represents the GMO communities' super node, the alternative health communities' super-node, the other communities' super node, and the non-Covid and non-vaccine conspiracy theory communities' super-node. Though the interactions all have a linear form with constant coupling coefficients g_R, g_B and $g_{G_i'}$, each of these interaction terms can be correctly thought of as the first term in a series expansion for a far more general functional form for the interaction. The sum over i could include the category $G(t)$ itself to reflect an intrinsic growth process, e.g. current members inciting friends and family who had not so far had an online presence.

We can write down similar equations to Eq. 1 for $R(t)$, $B(t)$ and all the separate $G_i'(t)$ terms for $i = 1, 2, \ldots$, hence yielding a set of n coupled first-order differential equations. We could also break the antis or pros into sub-categories like the neutrals, hence adding more sums to Eq. 1 and increasing the number and complexity of the resulting coupled differential equations. It would then be fascinating to explore how the system behaves dynamically at these different levels of aggregation, and determine what the best level of aggregation actually is. We leave that for future work, and instead we first obtain the simplest incarnation of the model in Eq. 1 for $R(t)$, $B(t)$ and $G(t)$, to illustrate the case of $n = 3$. Then we will show how we can analyze the case of any n exactly by assuming that all the non-$G(t)$ super-nodes are in steady state. This will give us an explicit expression for $G(t)$ for any n. We will then show that the resulting $G(t)$ behavior for any n exhibits a tipping point that is driven by the sign of the sum of the coupling terms, not their individual values. Then we will obtain estimates of this sum in two independent ways, and show they give remarkably consistent results. Then we will use this value for a prediction of the future of $G(t)$ based on an estimate of current conditions. Then we will discuss strategies to push the system back over the tipping point into safer territory.

To start, we analyze the illustrative case of three supernodes comprising all antis, all pros and all neutrals aggregated together, i.e. $n = 3$. Based on our observations of the actual empirical online content, we make the reasonable approximation that the pro communities are focused on emitting establishment guidance to the entire population, including of course the neutrals and antis. Hence they are not significantly influenced by the activity of the antis or the neutrals. So the equation for $B(t)$ is not coupled to the equations for the antis or the neutrals. By contrast, the anti communities are significantly influenced by this guidance emitted by the pro communities, in that they turn it into their own versions (including misinformation) and then feed it to the neutrals in order to raise the neutrals' concern about establishment guidance. They are not significantly influenced by the narratives of the neutrals themselves. Hence the equation for $R(t)$ is coupled to the equation for the pros. Finally, the neutral communities are significantly influenced by the guidance they receive from all sides: from the pros, the antis, and from other neutrals. Hence the equation for $G(t)$ is coupled to the equation for the pros and the antis. We then make the approximation of replacing the term $\sum_i g_{G_i'} G_i'$ in $\sum_i g_{G_i'} \left(G_i' - G\right)$ in Eq. 1 by a time-averaged version $g_G G_0$ and we also sum all the coupling terms, yielding $g_G \left(G_0 - G\right)$. The term G_0 could also include the impact of the category $G(t)$ itself, i.e. an intrinsic growth term for $G(t)$.

Adopting these reasonable approximations, the resulting equations for this $n = 3$ system dynamics become:

$$\begin{aligned} \frac{\partial R}{\partial t} &= r_R \left(R_0 - R\right) + r_B \left(B - R\right) \\ \frac{\partial B}{\partial t} &= b_B \left(B_0 - B\right) \\ \frac{\partial G}{\partial t} &= g_R \left(R - G\right) + g_B \left(B - G\right) + g_G \left(G_0 - G\right). \end{aligned} \quad (2)$$

Figure 6 right-hand panels show the results for two slightly different sets of the couplings in Eq. 2. While of course better fits can be achieved using AI tools. it is insightful when seeking an understanding of the system dynamics, to develop a transparent mathematical model that is minimal in terms of the number of parameters, and which can be easily interpreted, examined and understood.

We now turn to obtaining an equation for the behavior of $G(t)$ that we can then solve exactly. Given the longevity of the Covid-19 pandemic and the current 'steady-state' in terms of the world now having established vaccines, we start by assuming that the current activity levels of the pros and antis have reached a steady state, i.e. we take $R(t) \to R^*$ and $B(t) \to B^*$ which are constants. Hence from Eq. 2

$$\frac{\partial G}{\partial t} = g_R \left(R^* - G\right) + g_B \left(B^* - G\right) + g_G \left(G_0 - G\right). \quad (3)$$

The exact solution of Eq. 3 is then:

$$G(t) = G^* - \left[G^* - G\left(t = 0\right)\right] e^{-[g_R + g_B + g_G]t} \quad (4)$$

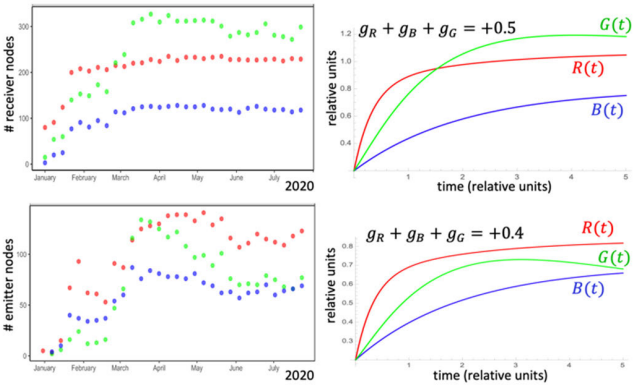


FIGURE 6. Comparison of empirical data (left panels) to the theoretical model output from Eq. 2 (right panels), for the period of maximal public uncertainty about Covid-19, i.e. 2020 prior to announcement of Covid-19 vaccines and hence the period with likely the richest dynamical system behavior. We do not seek a rigorous best fit, but just want to capture qualitatively the trends and hence estimate values for the sum $(g_R + g_B + g_G)$. Left panels show two possible empirical measures of $R(t)$, $B(t)$ and $G(t)$. Bottom left: Empirical data (circles) at each time t show the number of anti (red), pro (blue) and neutral (green) communities that at that timestep feature a piece of Covid-19 guidance and have a link into them from another community, hence making them emitters. Top left: Empirical data for the number of communities that have a link to a community that features a piece of Covid-19 guidance at that timestep, hence making them receivers. Right panels: output from Eq. 2 for $R(t)$, $B(t)$, $G(t)$. Model parameter values: top panel right $g_B = -1.0$; bottom panel right $g_B = -1.1$. For both panels, $r_B = 1.0$, $g_R = 0.5$, $r_R = 2.0$, $b_B = 0.5$, $g_G = 1.0$.

with G^* given by:

$$G^* = \frac{\left(g_R R^* + g_B B^* + g_G G_0\right)}{\left(g_R + g_B + g_G\right)}. \quad (5)$$

By adding the extra terms appearing in Eq. 1, we can easily generalize Eqs. 3-5 to provide an approximation for any number $n > 3$ and hence any number of neutral categories:

$$\frac{\partial G}{\partial t} = g_R \left(R^* - G\right) + g_B \left(B^* - G\right) + \sum_i g_{G_i'} \left(G_{i,0}' - G\right). \quad (6)$$

The exact solution of Eq. 6 for future times is then given by:

$$G(t) = G^* - \left[G^* - G\left(t = 0\right)\right] e^{-\left[g_R + g_B + \sum_i g_{G_i'}\right]t} \quad (7)$$

with G^* now given by:

$$G^* = \frac{\left(g_R R^* + g_B B^* + \sum_i g_{G_i'} G_{i,0}'\right)}{\left(g_R + g_B + \sum_i g_{G_i'}\right)}. \quad (8)$$

This expression for $G(t)$ contains a sum of the couplings. It does not matter where the individual couplings came from, nor how many there are. It is just the sum of the couplings that dictates the future values $G(t)$ in Eq. 7.

To predict the future behavior of $G(t)$, we therefore need to estimate current values for the sum of the couplings from the data. We now do this in two completely independent ways and show they give consistent results. First, we consider the simple case of $n = 3$, and specifically Eq. 2. The two left panels in Fig. 6 show the empirical data for two separate but

reasonable empirical measures of $R(t)$, $B(t)$, and $G(t)$. The bottom left-hand panel shows the empirical number of pro (blue), anti (red), neutral (green) communities that at each timestep t feature a piece of Covid-19 guidance and have a link into them from another community, hence making them active emitters at timestep t. Similarly, the top left-hand panel shows the empirical number of pro (blue), anti (red), neutral (green) communities that at each timestep t have a link to a community that features a piece of Covid-19 guidance, hence making them active receivers at timestep t. The right-hand panels show that the model curves from Eq. 2 capture the general empirical trends. To make these fits more demanding, we set most of the parameter values to be the same for each case, leaving the coupling values to differ just slightly. We then use the average and spread of the two numerical estimates for the sum of the couplings (+0.5 and +0.4) to obtain a single estimate of $(g_R + g_B + g_G) = 0.45 \pm 0.05$.

A second, and entirely independent, estimate of the sum of the couplings uses the super-node link weightings shown in Fig. 5a, and hence ties together the pure network analysis approach based on counting links with the purely dynamical description in the above equations. We start by bundling together the couplings from all the neutrals so that $(g_R + g_B + \sum_i g_{G_i'})$ becomes $(g_R + g_B + g_G)$. We know from our empirical observations of their content that when $B(t)$ becomes larger, the concern (online activity) of the neutrals and hence $G(t)$ becomes smaller, i.e. it makes the change of $G(t)$ in time more negative. This implies $g_B < 0$. By contrast, when $R(t)$ becomes larger, the concern (online activity) of the neutrals and hence $G(t)$ becomes larger, i.e. it makes the change of $G(t)$ in time more positive. This implies $g_R > 0$. Similarly, when the activity of other neutrals which we have approximated in an average way using the constant G_0 becomes larger, the concern (online activity) of the neutrals and hence $G(t)$ becomes larger, i.e. it makes the change of $G(t)$ in time more positive. This implies $g_G > 0$. We hence sum the link weights in Fig. 5a with these signs for antis (+75), pros (−19), alternative health (+13) plus the 80 links between parenting communities assuming an even split of ±40, i.e. $(+75) + (-19) + (+13) + (+40) + (-40) = +69$. Normalizing by the total $(75+19+13+40+40) = 187$ gives $(g_R + g_B + g_G) = +69/187 = 0.37$.

Given the fact that these two estimates of $(g_R + g_B + g_G)$ are both very crude and independent, it is pleasing that they have similar values (0.45 ± 0.05 compared to 0.37) and also have the same sign. It is also pleasing that the signs of the individual coupling terms that emerge from the fit in Fig. 6, are the same as those predicted based on our observations of the online content: i.e. $g_B < 0$, $g_R > 0$ and $g_G > 0$. These internal consistencies suggest that our model and its parameters are indeed interpretable and have sensible meanings.

We now want to use this estimate of $(g_R+g_B+g_G)$ to make a forecast for $G(t)$ into the future, based on a crude estimate of current conditions ($t = 0$). We stress again that rather than aiming to provide the most accurate forecast, we are looking to forecast trends in behaviors while also illustrating how the same methodology could be used for more detailed sets of equations. We will assume that the sum of the couplings does not change much over time and hence $(g_R + g_B + g_G) > 0$ remains true. It also seems reasonable to infer from reports in the media and the online comments that the current level of concern (i.e. level of online activity) of the mainstream parenting communities, has not yet reached its peak, i.e. $G(t = 0)$ is below its potential peak value G^* and hence $[G^* - G\,(t = 0)] > 0$ in Eq. 4. Since $(g_R + g_B + g_G) > 0$ in Eq. 4, this means $G(t)$ predicted by Eq. 4 will increase going forward in time toward G^* as shown in Fig. 5b.

Hence our simple yet exactly solvable model predicts that $G(t)$ will become endemic based on current conditions, i.e. the concern (online activity) of the mainstream parenting communities will increase going into the future. Specifically, it predicts that $G(t)$ will eventually tend toward some potentially high endemic level G^*. This is not good news of course.

If we could engineer the situation $(g_R + g_B + g_G) < 0$ instead, then $G(t)$ in Eq. 4 would decrease and become zero at some time in the near future, i.e. it would move the future behavior of the concern among mainstream parenting communities leftwards across the tipping point at $(g_R + g_B + g_G) = 0$ in Fig. 5b (thick black arrow). But making g_R or g_B more negative to achieve this, is likely too hard since it requires making the antis less concerning or the pros even more reassuring.

A direct reward from our analysis is that one could instead achieve $(g_R + g_B + g_G) < 0$ and hence $G(t) \to 0$, by making g_G more negative. This amounts to increasing the mainstream parenting communities' coupling to other types of neutral communities that seem unconcerned despite having seen the same online material [40] (Fig. 5a). This new strategy could be referred to as 'peer reassurance'. Using the above link analysis, the required reduction in g_G of >69 means that Facebook needs to encourage the creation of at least 69 new page links per 64 parenting communities to less concerned communities (e.g. pet-lover communities), i.e. a ratio of new links to pages of at least 1.1:1. These are of course crude estimates, but hopefully can help spark more detailed analyses and discussions. They certainly go beyond existing verbal guesswork.

This new strategy of increasing mainstream parenting communities' coupling to less concerned mainstream communities, has some additional advantages. First, it could help avoid contentious removal or censorship of content, users or communities, all of which are problematic. This is illustrated by the widely circulating conspiracy theory that Covid-19 vaccines contain tracking devices, hence allowing personal information to be read from foreheads: it turns out that this derives from a published article in a highly respected journal [41] that seems to indeed provide scientific proof-of-concept for such a technology. Hence moderator efforts to blanket label this as wrong science can backfire and inadvertently increase concern, as has already happened judging from the community

narratives that we see. Second, a much-publicized alternative approach of inoculating sets of communities within part of the machinery (Fig. 4a) could inadvertently result in some non-Covid/non-vaccine conspiracy theory communities ending up closer to mainstream parenting communities, as shown explicitly by our simulation in the SI where we mimic an intervention that temporarily decouples GMO communities within the network. Third, this new policy of increasing parenting communities' coupling to less concerned mainstream communities, can be operated in real-time and at scale since neither the nodes nor links tend to change on a daily basis. Any sudden shifts in node type could be captured by continually feeding community narratives into standard machine-learning tools. The SI demonstrates this, with the resulting word clouds and topic lists mimicking the manual classification types.

V. LIMITATIONS OF OUR STUDY

We have tried to point out our study's assumptions and approximations within the main text. Additional limitations are that there are many other social media platforms, apart from Facebook, that need to be explored. However, Facebook is the largest and furthermore we believe that similar behaviors will arise in any platform where communities can form. There is also the question of influence of external agents or entities [16]. However, these social media communities tend to police themselves for bot-like or troll behavior. It would of course be useful to see how our results apply at the next scale across all platforms. We hope to move beyond such limitations in future work.

VI. CONCLUSION

Our study goes beyond current approaches to misinformation that represent little more than verbal guesswork, by developing a mechanistic understanding of how misinformation manages to thrive in the online social media system. Our study is to our knowledge the first attempt to develop a generative, quantitative engineering-like theory to tackle misinformation and its dynamical evolution at scale, and hence also goes well beyond descriptive statistical studies and narratives about what has happened in the past. That is not how engineering is done, and we hope this paper helps break the current trend of discussing social media misinformation in this way. Specifically, our study revealed a strengthening of the bond between conspiracy theory communities that promote misinformation about climate change, fluoride, chemtrails and 5G, and mainstream parent communities—and showed how alternative health communities acted as critical conduits. Furthermore, it showed how an adjacent core of tightly bonded, anti-vaccination communities injects additional Covid-19 and vaccine-specific misinformation. We discussed how this provides a two-pronged machinery that can generate new pieces of misinformation without needing any new news. Our analysis also showed why Facebook's own scheme to supply reliable information about vaccines and Covid-19 was not efficient, and why targeting the largest communities does not work.

We then developed a simple, yet exactly solvable, dynamical equation of the type studied widely in engineering, that shows how tailoring the connectivity of mainstream communities can prevent them from tipping toward such misinformation. Our conclusions should be applicable to any social media platform that has in-built community features and hence provide the basis for a new engineering approach to solving online misinformation and harms at scale.

Future extension of this work could include sentiment analysis to explore differences in characteristics within and across categories of communities. Another interesting extension would be to try to repeat all the results using an entirely AI approach. Comparing to the present analysis would then give insight into possible new results that AI could add, and where new AI tools might be usefully developed. More broadly, the relevance of social media research to engineering is entirely consistent with IEEE's position as the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity, since social media is among the most important—but also potentially dangerous—of the new technologies impacting humanity.

ACKNOWLEDGMENT

The authors would like to thank Nicolas Velasquez for helping obtain the empirical data and graphics shown in Fig. 2 and Fig. 6 left panels, and Rashmi Menon for help with the graphics in Fig. 4b.

REFERENCES

[1] R. Brown. (2020). *Counteracting Dangerous Narratives in the Time of COVID-19 Over Zero*. [Online]. Available: https://projectoverzero.org/newsandpublications

[2] N. Calleja, A. AbdAllah, N. Abad, and N. Ahmed, "A public health research agenda for managing infodemics: Methods and results of the first WHO infodemiology conference," *JMIR Infodemiol.*, vol. 1, no. 1, 2021, Art. no. 30979, doi: 10.2196/30979.

[3] The UK Home Affairs Select Committee. (2016). *Hate Crime: Abuse, Hate and Extremism Online*. [Online]. Available: https://publications.parliament.U.K./pa/cm201617/cmselect/cmhaff/609/609.pdf

[4] A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Science vs conspiracy: Collective narratives in the age of misinformation," *PLoS ONE*, vol. 10, no. 2, Feb. 2015, Art. no. e0118093. [Online]. Available: https://journals.plos.org/plosone/article%3Fid%3D10.1371/journal.pone.0118093

[5] R. Diresta, *Virality and Viruses: The Anti-Vaccine Movement and Social Media*. Accessed: Nov. 8, 2018. [Online]. Available: https://nautilus.org/napsnet/napsnet-special-reports/of-virality-and-viruses-the-anti-vaccine-movement-and-social-media/

[6] H. Larson, "A lack of information can become misinformation," *Nature*, vol. 580, no. 7803, p. 306, 2020.

[7] N. F. Johnson, N. Velasquez, N. J. Restrepo, R. Leahy, N. Gabriel, S. El Oud, M. Zheng, P. Manrique, S. Wuchty, and Y. Lupu, "The online competition between pro- and anti-vaccination views," *Nature*, vol. 582, no. 7811, pp. 230–233, 2020, doi: 10.1038/s41586-020-2281-1.

[8] N. Velásquez, R. Leahy, N. Johnson Restrepo, Y. Lupu, R. Sear, N. Gabriel, O. K. Jha, B. Goldberg, and N. F. Johnson, "Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms," *Sci. Rep.*, vol. 11, no. 1, 2021, Art. no. 11549.

[9] B. Nogrady, "'I hope you die': How the COVID pandemic unleashed attacks on scientists," *Nature*, vol. 598, no. 7880, pp. 250–253, Oct. 2021.

[10] R. F. Sear, N. Velasquez, R. Leahy, N. Johnson Restrepo, S. El Oud, N. Gabriel, Y. Lupu, and N. F. Johnson, "Quantifying COVID-19 content in the online health opinion war using machine learning," *IEEE Access*, vol. 8, pp. 91886–91893, 2020, doi: 10.1109/ACCESS.2020.2993967.

[11] S. Kemp. (2021). *Hootsuite*. [Online]. Available: https://www.hootsuite.com/resources/digital-trends

[12] S. Frenkel, D. Alba, and R. Zhong, "*Surge of Virus Misinformation Stumps Facebook and Twitter*. New York, NY, USA: The New York Times, Mar. 8, 2020. [Online]. Available: https://www.nytimes.com/2020/03/08/technology/coronavirus- misinformation-social-media.html

[13] R. Y. Moon, A. Mathews, R. Oden, and R. Carlin, "Mothers' perceptions of the internet and social media as sources of parenting and health information: Qualitative study," *J. Med. Internet Res.*, vol. 21, no. 7, Jul. 2019, Art. no. e14289.

[14] T. Ammari and S. Schoenebeck, "'Thanks for your interest in our Facebook group, but it's only for dads' social roles of stay-at-home dads," in *Proc. CSCW*, San Francisco, CA, USA, Feb./Mar. 2016, pp. 1363–1375.

[15] R. Laws, A. D. Walsh, K. D. Hesketh, K. L. Downing, K. Kuswara, and K. J. Campbell, "Differences between mothers and fathers of young children in their use of the internet to support healthy family lifestyle behaviors: Cross-sectional study," *J. Med. Internet Res.*, vol. 21, no. 1, Jan. 2019, Art. no. e11454.

[16] D. Centola, J. Becker, D. Brackbill, and A. Baronchelli, "Experimental evidence for tipping points in social convention," *Science*, vol. 360, no. 6393, pp. 1116–1119, Jun. 2018.

[17] P. W. Anderson, "Antiferromagnetism theory of superexchange interaction," *Phys. Rev.*, vol. 79, no. 2, pp. 350–356, Jul. 1950, doi: 10.1103/PhysRev.79.350.

[18] F. H. Allport and M. Lepkin, "Wartime rumors of waste and special privilege: Why some people believe them," *J. Abnormal Social Psychol.*, vol. 40, no. 1, pp. 3–36, 1945.

[19] G. Pennycook and D. G. Rand, "Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking," *J. Personality*, vol. 88, no. 2, pp. 185–200, Apr. 2020.

[20] T. Burki, "Vaccine misinformation and social media," *Lancet Digit. Health*, vol. 1, no. 6, pp. e258–e259, Oct. 2019.

[21] A. J. Berinsky, "Rumors and health care reform: Experiments in political misinformation," *Brit. J. Political Sci.*, vol. 47, no. 2, pp. 241–262, 2017.

[22] N. Köbis and L. D. Mossink, "Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry," *Comput. Hum. Behav.*, vol. 114, Jan. 2021, Art. no. 106553, doi: 10.1016/j.chb.2020.106553.

[23] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.

[24] H. Ward, G. P. Garnett, K. H. Mayer, and G. A. Dallabetta, "Maximizing the impact of HIV prevention technologies in sub-Saharan Africa," *J. Int. AIDS Soc.*, vol. 22, no. S4, Jul. 2019, Art. no. e25319.

[25] H. Ward, *The U.K. has Record Death Tolls, Yet Still the Government has no Clear COVID Strategy*. London, U.K.: The Guardian, 2011. Accessed: Jan. 21, 2021. [Online]. Available: https://www.theguardian.com/commentisfree/2021/jan/21/U.K.-record-death-tolls-no-clear-covid-strategy

[26] R. Smith, S. Cubbon, and C. Wardle. (2020). *Under the Surface: COVID-19 Vaccine Narratives, Misinformation & Data Deficits on Social Media*. [Online]. Available: https://firstdraftnews.org/vaccine-narratives-report-summary-november-2020

[27] A. Gruzd and P. Mai, "Inoculating against an infodemic: A Canada-wide COVID-19 news, social media, and misinformation survey," May 2020. [Online]. Available: https://ssrn.com/abstract=3597462

[28] S. Lewandowsky. (2020). *The Debunking Handbook*. [Online]. Available: https://sks.to/db2020

[29] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, pp. 1146–1151, May 2018, doi: 10.1126/science.aap9559.

[30] J. Donovan, "Social-media companies must flatten the curve of misinformation," *Nature*, Apr. 2020, doi: 10.1038/d41586-020-01107-z.

[31] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter," *PLoS ONE*, vol. 6, no. 12, 2011, Art. no. 26752.

[32] J. P. Onnela, J. Saramäki, and J. Hyvönen, "Structure and tie strengths in mobile communication networks," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 18, pp. 7332–7336, 2007, doi: 10.1073/pnas.0610245104.

[33] A. Bechmann, "Tackling disinformation and infodemics demands media policy changes," *Digit. Journalism*, vol. 8, no. 6, pp. 855–863, Jul. 2020.

[34] C. A. Klofstad, J. E. Uscinski, J. M. Connolly, and J. P. West, "What drives people to believe in Zika conspiracy theories?" *Palgrave Commun.*, vol. 5, no. 1, pp. 1–8, Dec. 2019, doi: 10.1057/s41599-019-0243-8.

[35] A. M. Guess, B. Nyhan, and J. Reifler, "Exposure to untrustworthy websites in the 2016 U.S. election," *Nature Hum. Behav.*, vol. 4, no. 5, pp. 472–480, May 2020.

[36] T. A. Holroyd, A. C. Howa, P. L. Delamater, N. P. Klein, A. M. Buttenheim, R. J. Limaye, T. M. Proveaux, S. B. Omer, and D. A. Salmon, "Parental vaccine attitudes, beliefs, and practices: Initial evidence in California after a vaccine policy change," *Hum. Vaccines Immunotherapeutics*, vol. 17, no. 6, pp. 1675–1680, Jun. 2021.

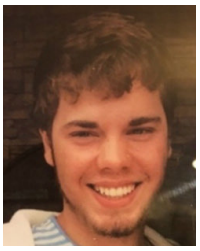
[37] N. Velásquez, P. Manrique, R. Sear, R. Leahy, N. J. Restrepo, L. Illari, Y. Lupu, and N. F. Johnson, "Hidden order across online extremist movements can be disrupted by nudging collective chemistry," *Sci. Rep.*, vol. 11, no. 1, 2021, Art. no. 9965.

[38] P. D. Manrique, M. Zheng, Z. Cao, E. M. Restrepo, and N. F. Johnson, "Generalized gelation theory describes onset of online extremist support," *Physical. Rev. Lett.*, vol. 121, no. 4, 2018, Art. no. 048301.

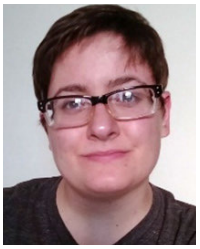
[39] N. F. Johnson, M. Zheng, Y. Vorobyeva, A. Gabriel, H. Qi, N. Velasquez, P. Manrique, D. Johnson, E. Restrepo, C. Song, and S. Wuchty, "New online ecology of adversarial aggregates," *Science*, vol. 352, no. 6292, pp. 1459–1463, 2016.

[40] M. J. Gelfand, J. R. Harrington, and J. C. Jackson, "The strength of social norms across human groups," *Perspect. Psychol. Sci.*, vol. 12, no. 5, pp. 800–809, Sep. 2017, doi: 10.1177/1745691617708631.

[41] K. J. McHugh, L. Jing, S. Severt, and M. Cruz, "Biocompatible near-infrared quantum dots delivered to the skin by microneedle patches record vaccination," *Sci. Transl. Med.*, vol. 11, no. 523, Dec. 2019, Art. no. eaay7162, doi: 10.1126/scitranslmed.aay7162.



NICHOLAS J. RESTREPO received the B.Sc. degree in business administration from Carnegie Mellon University, in 2018. He is currently a Founding Partner at ClustrX LLC. He has business experience in small and large companies in Europe and South America.



LUCIA ILLARI received the B.A. degree in physics from the Barnard College, Columbia University. They are currently pursuing the Ph.D. degree in physics with George Washington University. Their Ph.D. research is on complex adaptive systems, focusing on the mathematical, physical and data analysis of dynamical network systems.

RHYS LEAHY received the B.A. degree in international studies from American University. She is currently a Research Scientist with George Washington University, and a Founding Partner at ClustrX LLC. She was awarded scholarships from the Department of State to pursue advanced language studies in Russia and Tajikistan.

RICHARD F. SEAR received the B.Sc. degree with a major in computer science and minors in mathematics and physics from George Washington University. He is currently a Research Scientist at George Washington University. He has worked on emotion recognition neural networks for Buchanan & Edwards, and NER and topic recognition models as part of the Johns Hopkins SCALE program.

YONATAN LUPU received the B.A. and J.D. degrees from Georgetown University, and the M.A. and Ph.D. degrees from the University of California at San Diego. He is currently a Professor with the Department of Political Science, George Washington University. His current research interests include political violence, digital authoritarianism, human rights abuses, and violent online extremism.

NEIL F. JOHNSON received the B.A. and M.A. degrees from Cambridge University, U.K., and the Ph.D. degree from Harvard University as a Kennedy Scholar. He is currently a Professor with the Physics Department, George Washington University. His research interests include complex systems and networks. He is a fellow of the American Physical Society. He was a recipient of the 2018 Burton Award from the APS.