IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version April 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Quantifying COVID-19 content in the online health opinion war using machine learning

R.F. Sear[1], N. Velásquez[2,3], R. Leahy[2,4], N. Johnson Restrepo[2,4], S. El Oud[5], N. Gabriel[5], Y. Lupu[6] and N.F. Johnson[5,2]

[1]Department of Computer Science, George Washington University, Washington D.C. 20052 USA
[2]Institute for Data, Democracy and Politics, George Washington University, Washington D.C. 20052 USA
[3]Elliott School of International Affairs, George Washington University, Washington D.C. 20052 USA
[4]ClustrX LLC, Washington D.C. 20007 USA
[5]Department of Physics, George Washington University, Washington D.C. 20052 USA
[6]Department of Political Science, George Washington University, Washington D.C. 20052 USA

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

ABSTRACT A huge amount of potentially dangerous COVID-19 misinformation is appearing online. Here we use machine learning to quantify COVID-19 content among online opponents of establishment health guidance, in particular vaccinations ("anti-vax"). We find that the anti-vax community is developing a less focused debate around COVID-19 than its counterpart, the pro-vaccination ("pro-vax") community. However, the anti-vax community exhibits a broader range of "flavors" of COVID-19 topics, and hence can appeal to a broader cross-section of individuals seeking COVID-19 guidance online, e.g. individuals wary of a mandatory fast-tracked COVID-19 vaccine or those seeking alternative remedies. Hence the anti-vax community looks better positioned to attract fresh support going forward than the pro-vax community. This is concerning since a widespread lack of adoption of a COVID-19 vaccine will mean the world falls short of providing herd immunity, leaving countries open to future COVID-19 resurgences. We provide a mechanistic model that interprets these results and could help in assessing the likely efficacy of intervention strategies. Our approach is scalable and hence tackles the urgent problem facing social media platforms of having to analyze huge volumes of online health misinformation and disinformation.

INDEX TERMS COVID-19, machine learning, topic modeling, mechanistic model, social computing

I. INTRODUCTION

Scientific experts agree that defeating COVID-19 will depend on developing a vaccine. However, this assumes that a sufficiently large proportion of people would receive a vaccine so that herd immunity is achieved. Because vaccines tend to be less effective in older people, this will require younger generations to have very high COVID-19 vaccination rates in order to guarantee herd immunity [1]. Yet there is already significant opposition to existing vaccinations, e.g. against measles, with some parents already refusing to vaccinate their children. Such vaccine opposition increased the number of cases in the 2019 measles outbreak in the U.S. and beyond [2]. Any future COVID-19 vaccine will likely face similar opposition [3][4]. Mandatory COVID-19 vaccinations for schoolchildren could trigger a global public health conflict. A better understanding of such opposition ahead of a COVID-19 vaccine is therefore critical for scientists, public health practitioners, and governments.

Online social media platforms, and in particular the built-in communities that platforms like Facebook (FB) feature, have become popular fora for vaccine opponents (anti-vax) to congregate and share health (mis)information. Such misinformation can endanger public health and individual safety [1][4]. Likewise, vaccine supporters (pro-vax) also congregate in such online communities to discuss and advocate for professional public health guidance. Well before COVID-19, there was already an intense online conflict featuring anti-vax communities and pro-vax communities. Within anti-vax communities, the narratives typically draw on and generate misinformation about establishment medical guidance and distrust of the government, pharmaceutical industry, and new technologies such as 5G communications [1][4][5]. Adding fuel to this fire, the January 2020 birth of the COVID-19 "infodemic" has led to a plethora of misinformation in social media surrounding COVID-19 that directly threatens lives [6]. For example, harmful "cures" are being proposed such as drinking fish tank additives, bleach, or cow urine, along with coordinated threats against public health officials like Dr. Anthony Fauci, director of the U.S. National Institute of Allergic and Infectious Diseases [7]. Moreover, false rumors have been circulating that individuals with dark skin are immune to COVID-19. These may have contributed to more relaxed social distancing among some minorities and hence

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993967, IEEE Access

their over-representation as victims. In Chicago and Louisiana as of early April 2020, ~70% of the fatalities were African Americans even though this demographic only makes up ~30% of the population [8] [9]. In addition, the world has witnessed an alarming rise in COVID-19 weaponization against the Asian community [10][11][12]. It is also clear that such misinformation is not a fringe phenomenon, and can instead be very widely held as true within the general population. Indeed, a recent Pew study [13] found that ~30% of Americans believe the COVID-19 virus was likely created in a laboratory, despite statements from infectious disease experts to the contrary.

Unfortunately, the sheer volume of new online content and the speed with which it spreads, means that social media companies are struggling to contain such health misinformation [14][15]. Making matters worse, people around the world are spending more time on social media due to social distancing imposed during the COVID-19 pandemic. This increases the likelihood that they become exposed to such misinformation, and as a result they may put themselves and their contacts at risk with dangerous COVID-19 remedies, cures and falsehoods.

The present study is motivated by both these needs: (1) the need for a deeper understanding of this intersection between online vaccination opposition and the online conversation surrounding COVID-19; and (2) the need for an automated approach since the sheer volume of new online material every day makes manual analysis a non-viable option going forward. We pursue an automated, machine learning approach that avoids the scalability limitations of manual content analysis. While the present paper is just the first step in a challenging longer-term goal, the automated approach that we present allows the following questions to be addressed: How did COVID-19 change the online conversation within anti-vaccination and pro-vaccination communities over the two month period in early 2020 when the disease became a global threat; and what do the topical changes that we observe in the anti-vax and pro-vax online communities' narratives, imply about their relative abilities to attract new supporters going forward?

Unlike many existing works, this study does not use Twitter data [16][17] since it is known that Twitter is more of a broadcast medium for individual shout-outs, whereas discussions and narratives tend to be nurtured in in-built online community spaces that are a specific feature of platforms like Facebook (e.g., fan page) [18]. Twitter does not have such in-built community spaces. In the present methodology, generalized from Refs. [19] and [20], data is collected from these online communities, specifically Facebook Pages that support either anti-vaccination or pro-vaccination views. This information is publicly available and does not require any individual details, thereby avoiding any privacy concerns – just as understanding the content of a conversation among a crowd of people in an open, real-world public space does not require knowledge of any personal details about the individuals within that crowd. Details of our approach are given in Sec. II and the Appendix. A third difference between this study and previous ones is that the machine learning findings here are interpreted in terms of a mechanistic model (Sec. IV) that captures the general trend for the coherence in the online conversations over time. Though much work still needs to be done, this study therefore provides a first step toward a fully automated but interpretable understanding of the growing public health debate concerning vaccines and COVID-19.

II. DATA AND MACHINE LEARNING ANALYSIS

The terms 'Facebook Page' and 'cluster' are used interchangeably here since each Facebook Page is a cluster of people. Facebook Pages, also known as fan pages or public pages, are accounts that represent organizations, causes, communities, or public figures. According to Facebook's policies, "Content posted to a Page is public and can be viewed by everyone who can see the Page" [see Ref. 21, Sec. 5]. A Facebook Page is different from a Facebook personal account. Personal accounts represent private individuals, and their posts and interactions are considered more private and targeted to their immediate contacts. This paper does not analyze data from personal accounts. Our methodology follows Refs. [19] and [20] by analyzing the public content of Facebook Pages for both anti-vaccination ("anti-vax") and pro-vaccination ("pro-vax") communities. The publicly available content of these online communities is obtained using a snowball approach, starting with a seed of manually identified pages discussing either vaccines, public policies about vaccination, or the pro-vs-anti vaccination debate. Then their connections to other fan pages are indexed. At each step, new clusters are evaluated through a combination of human coding and computer-assisted filters. To classify a cluster as being (1) anti-vax or pro-vax and (2) including COVID-19 content or not, we reviewed its posts and the Page's "about" section. Pro-vax and anti-vax classifications required that either (a) at least 2 of the most recent 25 posts dealt with the pro-vax or anti-vax debate, or (b) the page's title or "about" section described it as pro-vax or anti-vax. At least two researchers classified each cluster independently. If they disagreed on their suggested classification, a third researcher reviewed the posts and then all three reviewers discussed these cases. Agreement was reached in each case. This also enabled us to distinguish between content that is intended to be serious versus merely satirical. The self-weeding tendency within Facebook Pages tends to reduce material from bots and 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993967, IEEE Access

The content of these clusters was then bundled together separately for the anti-vax community and the pro-vax community, and the two resulting sets of content were analyzed using machine learning. Specifically, we used an unsupervised machine learning technique called Latent Dirichlet Allocation (LDA) [22] to analyze the emergence and evolution of topics around COVID-19. The LDA method models documents as distributions of topics and topics as distributions of words. During its training process, these distributions are adjusted to fit the dataset. The LDA method is described correctly in Wikipedia as [23] "[quote] .. a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. LDA is an example of a topic model and belongs to the machine learning toolbox and in wider sense to the artificial intelligence toolbox."

The coherence score provides a quantitative method for measuring the alignment of the words within an identified topic (see Ref. [22]). It is generated from a separate algorithm which is run over a trained LDA model. The overall coherence score of a single model is the arithmetic mean of its per-topic coherences. There are many different coherence metrics to evaluate per-topic coherence. We use C_V which is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized point-wise mutual information and the cosine similarity. It comprises collections of probability measures on how often top words in topics co-occur with each other in examples of the topics. We refer to Ref. [22] for a full explanation and discussion of C_V.

Machine learning automation can, in principle, help address the significant issues facing social media platforms by mechanically picking out material that requires attention from the huge haystack of online content. While this could help to better curtail online misinformation, one might rightly ask about its accuracy and reliability as compared to human analysts. This has been recently addressed in Ref. [24]. We use the same coherence metric (C_V) as these authors. They addressed the problem that topic models had previously given no guarantee on the interpretability of their output. Specifically, they produced several benchmark datasets with human judgements of the interpretability of topics and they found results that outperformed existing measures with respect to correlation to human ratings. They achieved this by evaluating 237,912 coherence measures on 6 different benchmarks for topic coherence, making this the biggest study of topic coherences at that time. Separately, we have done our own comparison for the general area of online hate and have found comparable consistency.

In summary, our machine learning approach identifies topics in the online narratives with high coherence, meaning the word groupings identified are strongly related according to the coherence scoring approach discussed earlier. Our human inspection of the word distribution making up each grouping showed that they do indeed correspond to reasonably distinct conversation topics. Details and examples are given in the Appendix.

III. RESULTS

The main focus here is in the endogenous development of COVID-19 conversation at the beginning of the global pandemic and prior to the first officially reported U.S. COVID-19 death on February 29, 2020 [25]. Hence we collected Facebook public post data for the period 1/17/2020-2/28/2020 inclusive. To assess the change over time, this period was divided into time intervals. Since having more time intervals would mean smaller amounts of data within each and hence more fluctuations, and since we are just interested in the change over time, two intervals were chosen of equal duration, T_1 and T_2. The first time-interval 1/17/2020-2/7/2020 (T_1) contains 774 total pro-vax posts and replies, and 3630 total anti-vax posts and replies. The second time-interval 2/7/2020-2/28/2020 (T_2) contains 673 total pro-vax posts and replies, and 3200 total anti-vax posts and replies. Hence our two equal time windows contains similar amounts of data. We checked that our results are relatively robust to other choices of time interval. Interestingly, T_1 roughly corresponds to the time when COVID-19 was largely seen as a problem in Asia, while T_2 roughly corresponds to the time during which it became a serious problem in Europe. For further reassurance that our data was representative of the COVID-19 conversation during these intervals, we also checked that the data split is similar to that for mentions of COVID-19 in article counts from worldwide anglophone newspapers and worldwide Google trends.

The LDA models were trained over posts in the following distinct groups: anti-vaccination posts in T_1, anti-vaccination posts in T_2, pro-vaccination posts in T_1, and pro-vaccination posts in T_2. For each of these sets, 10 separate LDA models were trained with the *number of topics* parameter ranging from 3-20, for a total of 180 models in each of the four groups. Fuller details are given in the Appendix. The C_V coherence algorithm was then run over each of these models and the coherence scores were averaged for each number of topics. These averaged scores are plotted in Figures 1B and 1C. Figure 1A shows the result of the same procedure applied to all posts in our dataset, and to all anti-vaccination posts, and to all pro-vaccination posts.

The coherence score C_V for the entire period of study (i.e. T_1+T_2) in Fig. 1A, is consistently larger across the number of topics for pro-vax than for anti-vax, suggesting that the pro-vax community overall has a more focused discussion around COVID-19 than the anti-vax. This is consistent with the pro-vax community featuring a more monolithic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993967, IEEE Access

discussion around public health -- namely, it is focused on advising people to follow professional medical guidance.

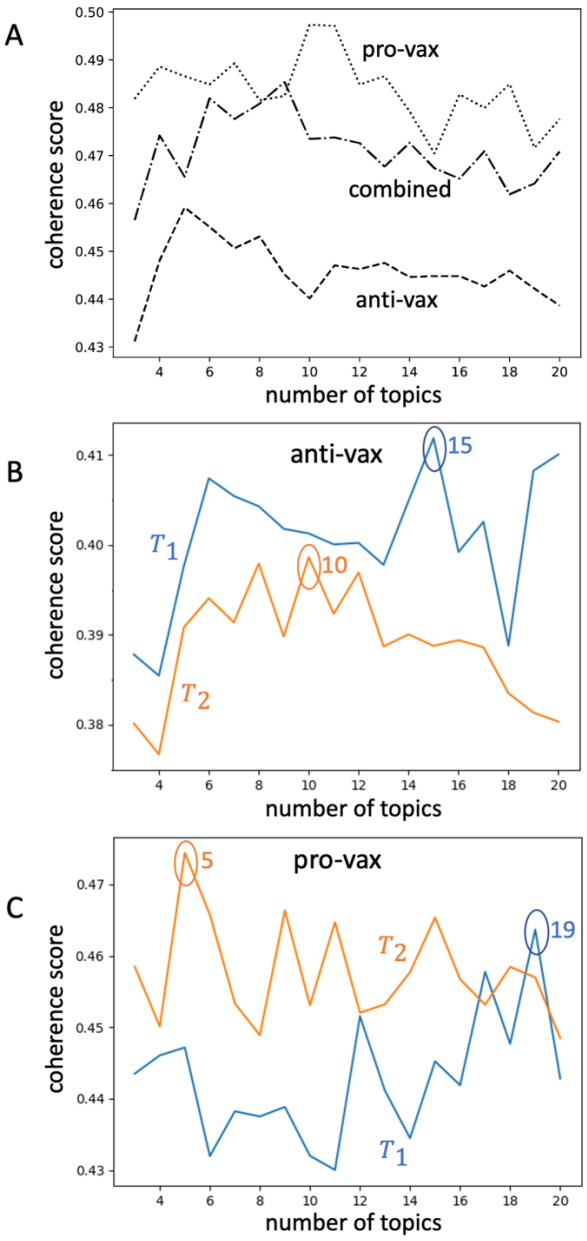

Fig. 1: Coherence scores C_V for (A) anti-vax (dashed line), pro-vax content (dotted line), and anti-vax combined with pro-vax (dashed-dotted line), calculated over the entire time period of study (T_1+T_2). (B) Anti-vax content for the separate time periods T_1 (blue line) and T_2 (orange line). The number of topics for which the coherence score C_V is a maximum is indicated, i.e. the optimal number of topics. The optimal number of topics for anti-vax moves from 15 to 10 from T_1 to T_2. (C) Pro-vax content for the separate time periods T_1 (blue line) and T_2 (orange line). The optimal number of topics for pro-vax moves from 19 to 5 from T_1 to T_2.

The bad news for the pro-vax community from this higher overall coherence, is that it is less well positioned to engage with the wide variety of more blurry, and often more extreme, COVID-19 narratives that are now circulating online. This represents a significant potential disadvantage for the pro-vax community in that it may therefore be less able to attract the attention of the many different types of users who are now entering this online space in search of a particular nuanced 'flavor' of COVID-19 narrative that appeals to them. These users could consequently be pulled toward the anti-vax cause.

Figures 1B and 1C indicate the change over time by comparing the curves of the coherence score across number of topics, for time periods T_1 and then T_2. The curve moves up from T_1 to T_2 for the pro-vax community (Fig. 1C) and the optimal number of topics shows a dramatic decrease from 19 to 5. This is consistent with the notion that the pro-vax community is working toward a common COVID-19 interpretation and narrative with fewer 'flavors' of discussion and interpretation than the anti-vax community. Again, while this may sound like a strength, it suggests that the pro-vax community overall is actually becoming *less* appealing over time to the many different types of new users who are in search of their own COVID-19 narrative 'flavor'. By contrast, the curves for the anti-vax community from T_1 to T_2 (Fig. 1B) show a far smaller reduction in the optimal number of topics (15 to 10) and the curves move down, in the opposite direction to the pro-vax. Hence the anti-vax compensates a small increase in focus (reduction in the optimal number of topics) with an overall reduction in coherence, i.e. these 10 topics for T_2 are effectively more blurry than the original 15 for T_1, and hence the overall anti-vax community is becoming more accommodating to the diverse population of new additions coming into the online health space over time.

Figure 2 shows a visualization with more detail about the information structure of the individual topics, and how far these topics are from one another in terms of informational distance. The plot is obtained using the pyLDAvis package [26] which provides a global view of the topics and how they differ from each other, while at the same time allowing for a deeper inspection of the terms most highly associated with each individual topic. This provides a novel method for implying the relevance of a term to a topic. The study in Ref. [26] showed that ranking terms purely by their probability under a topic, by contrast, is suboptimal for topic interpretation. We refer to Ref. [26] for full details of LDAvis.

The change in the pro-vax community from time period T_1 (Fig. 2C) to T_2 (Fig. 2D) is such that the optimal number of topics decreases (i.e., the number of circles decreases from 19 to 5 following Fig. 1C) and the topics evolve to become located mostly in the same portion of the space (i.e., toward the right-hand side of Fig. 2D). Following Fig. 1B, the change

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993967, IEEE Access

in the anti-vax community from time period T_1 (Fig. 2A) to T_2 (Fig. 2B) is such that the optimal number of topics starts off slightly smaller than the pro-vax, but although it also decreases over time (i.e., the number of circles decreases) there are more topics (i.e., more circles in Fig. 2B) than for the pro-vax in time period T_2 (Fig. 2D). Also, the topics seem more spread out across the space in Fig. 2B as compared to Fig. 2D. These observations are consistent with our earlier interpretations that the pro-vax community is more focused (equivalently, narrower) than the anti-vax community in terms of COVID-19 narratives, and that the pro-vax community is evolving toward a common COVID-19 interpretation and narrative with a lower diversity on offer than the anti-vax community.

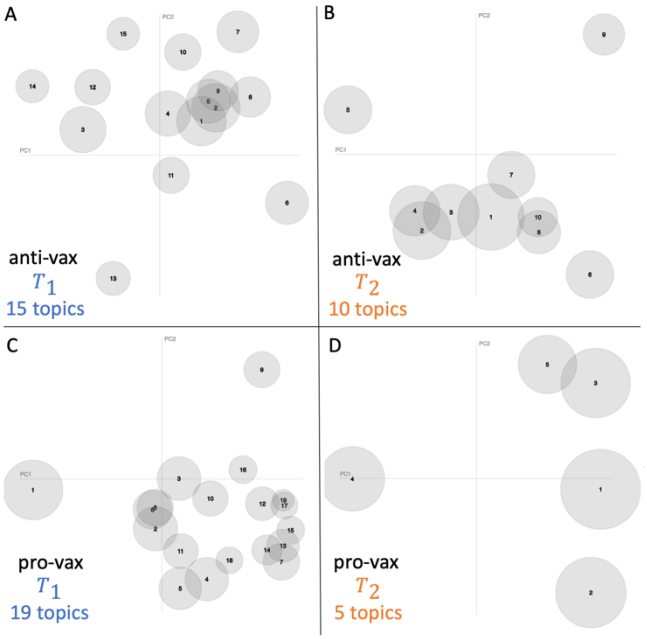

Fig. 2: Visualization of the informational structure of the individual topics, and how they relate to each other. This plot is obtained using pyLDAvis. The circles in each plot are the topics from Fig. 1 for which the average coherence score is highest, i.e. the optimal number of topics. Their size indicates the marginal topic distribution as discussed in detail in Ref. [26]**, while the two axes are principal components in the distribution analysis.**

IV. TOWARD A MECHANISTIC MODEL INTERPRETATION

We created a mechanistic model that further supports these empirical findings and provides a microscopic interpretation of the machine learning output. Specifically, we generated a computer simulation of an ecology of online components of the overall community content, each of which is characterized by a vector $\boldsymbol{x}=(x_1, x_2, \ldots)$ in which each component x_i signifies the strength of a given factor surrounding the online health debate, e.g. government control. The exact nature of these components does not need to be specified, i.e. whether they are words or short phrases. It just matters that there is a diverse ecology of such building blocks. This mechanistic model setup, while seemingly very simplistic, does indeed reflect the empirical observations and literature surrounding the themes of online discussions of vaccination opposition, as listed and studied in detail by Kata in Ref. [1]. We then carry out a simulation whereby these components are selected randomly to build up content. Components cluster together (or their clusters cluster together, if they are already in a cluster) if their $\boldsymbol{x}$ values are sufficiently similar (i.e. homophily in Fig. 3A) or different (i.e. heterophily in Fig. 3B). To illustrate the output of our model, Fig. 3 shows a one-dimensional version. We checked that a two-dimensional version gives similar results, though it is visually more complicated because of having the time component along the third dimension. Most importantly, it produces plots that are visually similar to those in Fig. 2. As can be seen from Fig. 3, the case of homophily (which is akin to building a more monolithic topic discussion with few flavors, like the pro-vax community) has a convergence that is quicker, as observed in Figs. 1 and 2 for the pro-vax community. By contrast, the case of heterophily (which is akin to building diverse topic discussions with many flavors, like the anti-vax community) is slower to gel, which is consistent with the anti-vax community in Figs. 1 and 2. The red dotted horizontal line in Figs. 3A and 3B gives an indication of the stage in the simulation that is broadly consistent with Figs. 2D and 2B for the pro-vax and anti-vax communities respectively.

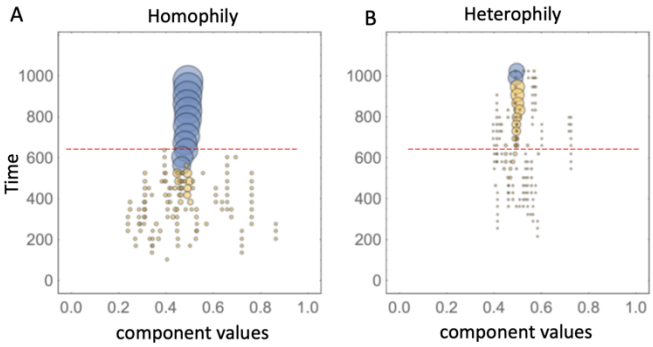

Fig. 3: Output from our mechanistic model in which clusters form if the component x-values are sufficiently similar (i.e., homophily in panel A) or different (i.e., heterophily in panel B).

The delay in the gelation time observed in Fig. 3B for heterophily (anti-vax) as compared to homophily (pro-vax) in Fig. 3A, can be derived analytically using mathematical analysis from statistical physics (see Ref. [27] for full details). In particular, we have been able to show that the time at which gelation emerges depends inversely on the average probability that two randomly picked components join the same cluster, which is smaller for heterophily than homophily and hence the gelation time is later for heterophily than homophily -- exactly as observed in Fig. 3. Similarly, it can be shown mathematically that the gelation sizes (akin to the sizes of the circles in Fig. 2) will be smaller for heterophily than homophily, as also observed in Fig 3.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993967, IEEE Access

Again, instead of this being good news for the pro-vax community, the simulation of this mechanistic model shows that the case of homophily (pro-vax) is less able to absorb an influx in new users with a range of $\boldsymbol{x}$ values, as compared to the case of heterophily (anti-vax). This is consistent with the idea stated earlier, that the anti-vax community appears more engaging to new users (e.g. parents with children of school-age who are wary of school vaccine requirements, or who fear government control) and hence the anti-vax will be more able to gain new supporters in the long run than the pro-vax.

V. LIMITATIONS OF THE STUDY

There are of course many limitations of this study. There are other social media platforms, apart from Facebook, that should be explored -- but Facebook is the largest. Similar behaviors should arise in any platform where communities can form. It will also be interesting, for example, to compare our findings to other studies focused on Twitter, where messaging is more in the form of short, individual statements [17]. 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. Further analysis is required of the details of the content. This will require going beyond just text and perhaps beyond LDA, since memes and images are also shared. Also, the generative model output needs to be compared in detail to the time-evolution of topics. Further research is also required to formulate the results across all platforms into detailed, actionable consequences for policy makers. These limitations will be addressed in future work.

VI. CONCLUSION

These findings suggest that the online anti-vax community is developing a more diverse and hence more broadly accommodating discussion around COVID-19 than the pro-vax community. As a result, the pro-vax community runs the risk of making itself less engaging to the heterogeneous ecology of potential new users who join the online COVID-19 discussion, and who may arrive online with a broad set of concerns, questions, and possibly preconceived notions, misinformation and even falsehoods.

The analysis in this paper also provides a first step toward eventually either replacing, or at least supplementing, the non-scalable efforts of human moderators tasked with identifying online misinformation. In addition, the mechanistic model (Fig. 3) could be used for what-if scenario testing of how quickly coherence develops and what the impact would be of breaking up the coherence around certain topics, e.g. by counter-messaging against individuals ingesting bleach or the even newer 'COVID Organics' that are circulating as a cure in Madagscar, Africa and beyond. This can be achieved by using the empirical analysis in Fig. 2 -- repeated over multiple consecutive time intervals -- to identify the growth of topics around new words which may be gaining popularity as a home cure (e.g. "bleach"). Then Facebook, for example, could post ads that specifically target these specific new words and topics, rather than blanket vanilla messaging promoting establishment medical science narratives.

Overall, this approach shows that a machine-learning algorithm, the LDA algorithm, identifies plausible topics within collections of posts from online communities surrounding the vaccine and COVID-19 debate. In addition to being able to handle large quantities of data, its results emerge quickly using statistical grouping techniques, instead of having to rely on potentially biased, slow and costly human labeling.

APPENDIX

As mentioned in the main text, the methodology starts with a seed of manually identified Facebook Pages discussing either vaccines, public policies about vaccination, or the pro-vs-anti vaccination debate. Then their connections to other fan pages are indexed. At each step, new findings are vetted through a combination of human coding and computer assisted filters. This snowball process is continued, noting that new links can often lead back to members already in the list and hence some form of closure can in principle be achieved. This process leads to a set containing many hundreds of pages for both the anti-vax and pro-vax communities. Before training the LDA models, several steps are employed to clean the content of these pages in a similar way to other LDA analyses in the literature:

Step 1: Mentions of URL shorteners are removed, such as "bit.ly" since these are fragments output by Facebook's CrowdTangle API.

Step 2: Many of the posts link to external websites. The fact that these specific websites were mentioned could itself be an interesting component of the COVID-19 conversation. Hence instead of removing them completely, the pieces ".gov", ".com", and ".org" were replaced with "__gov", "__com", and "__org", respectively. This operation effectively concatenates domains into a form that will not be filtered out by the later preprocessing steps.

Step 3: The posts are then run through Gensim's simple_preprocess function, which tokenizes the post on spaces and removes tokens that are only 1 or 2 characters long. This step also removes numeric and punctuation characters.

Step 4: Tokens that are in Gensim's list of stopwords, are removed. For example, "the" is not a good indication of a topic.

Step 5: Tokens are lemmatized using the WordNetLemmatizer from the Natural Language Toolkit NLTK, which converts all words to singular form and/or present tense.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993967, IEEE Access

Step 6: Tokens are stemmed using the SnowballStemmer from NLTK, which removes affixes on words.

Step 7: Any remaining fragments of URLs (other than domain) that are left over after stemming, such as “http” and “www”, are removed.

Steps 5 and 6 help ensure that words are compared fairly during the training process, and that if a particular word is a strong indicator of a topic, its signal is not lost just because it is used in many different forms. These steps rely on words existing in NLTK’s pretrained vocabulary. Any word not in this vocabulary is left unchanged. After this preprocessing, we then train the LDA models on the cleaned data. Specifically, 10 separate LDA models were trained with the “number of topics” parameter ranging from 3-20, for a total of 180 models in each of the two time intervals T_1 and T_2. The C_V coherence algorithm was then run over each of these models and the coherence scores were then averaged for each number of topics. To produce the results, multiple trials were run for each number of topics to ensure that the coherence for a particular number of topics is representative of what LDA models tend to find (and by extension a better fit for the data) and is not the result of unaccounted noise swaying the model to overfit in one way or another. These trials are independent because the random number generator for each LDA model was initialized with a different seed, ensuring that statistical inferences were not be repeated. The GitHub link is: https://github.com/searri/social-clustering-research/wiki/Coronavirus-Vax

The following illustrates the topic output, focusing on anti-vax in time interval T_2 in Fig. 2B. In this, 9 of the 10 topics had the word 'coronavirus' among the 5 highest weighted words in the topic; 4 were focused around 'coronavirus' and 'vaccine' co-occuring together. Others had 'vitamin', 'fear' and 'ddt' in relation to alternative treatments, and 'weapon' related to conspiracy theories of COVID-19's origin. Within one of the topics, which is focused around alternative health explanations and cures with words like 'vitamin' etc., illustrative posts include the following from Feb 8, 2020 in one of the 'Coalition for Vaccine Choice' pages, with spelling mistakes left as is: "The story of this FAKE "epidemic" with the "corona virus" from China is a cover-up story for the grim reality of the health problems due to 5G technology exposure coroborated with a lot of other factors: vaccination, poor alimentation in vitamins, bad water, air pollution, lack of sleep, etc.... scientists have shown that low level microwave EMF exposure can result in VGCC activation and elevated intracellular calcium". Meanwhile, for a topic focused on conspiracy theories with words such as 'weapon' and 'fear', an example phrase from a posting is "..keeping the world under the thumb of tyrants! You are soldiers, and that means that you are expendable by your trained nature. You are being micro-managed by people that give not one caring thought of you, five thousand miles away, that know little of the true nature of the battle". This illustrates the type of detailed analyses that we carried out to check our automated approach, and which underlie our claim that the groupings do correspond to reasonably distinct conversation topics.

ACKNOWLEDGMENT

CrowdTangle data are made available to the Institute for Data, Democracy, and Politics, via a grant from the John S. and James L. Knight Foundation.

REFERENCES

[1] A. Kata, “A postmodern Pandora’s box: Anti-vaccination misinformation on the Internet,” *Vaccine*, vol. 28, no. 7, pp. 1709–1716, Feb. 2010, doi: 10.1016/j.vaccine.2009.12.022.

[2] L. Givetash, “Global measles cases surge amid stagnating vaccinations,” *NBC News*, 2019. [Online]. Available: https://www.nbcnews.com/news/world/global-measles-cases-surge-amid-decline-vaccinations-n1096921. [Accessed: 13-Apr-2020].

[3] B. Martin, “Texas anti-vaxxers fear mandatory COVID-19 vaccines more than the virus itself,” *Texas Monthly*, 2020. [Online]. Available: https://www.texasmonthly.com/news/texas-anti-vaxxers-fear-mandatory-coronavirus-vaccines/.

[4] H. Larson, "Blocking information on COVID-19 can fuel the spread of misinformation," *Nature* vol. 580, no. 306, March 2020, doi: 10.1038/d41586-020-00920-w

[5] R. Schraer and E. Lawrie, “Coronavirus: scientists brand 5G claims ‘complete rubbish,’” *BBC News*, 2020. [Online]. Available: https://www.bbc.com/news/52168096. [Accessed: 05-Apr-2020].

[6] World Health Organization, “Coronavirus disease (COVID-19) advice for the public: myth busters,” *World Health Organization*, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters. [Accessed: 13-Apr-2020].

[7] K. Benner and M. Shear, “After threats, Anthony Fauci to receive enhanced personal security,” *The New York Times*, 2020. [Online]. Available: https://www.nytimes.com/2020/04/01/us/politics/coronavirus-fauci-security.html. [Accessed: 02-Apr-2020].

[8] S. Almasy, H. Yan, and M. Holcombe, “Coronavirus pandemic hitting some African-American communities extremely hard,” *CNN Health*, 2020. [Online]. Available: https://www.cnn.com/2020/04/06/health/us-coronavirus-updates-monday/index.html. [Accessed: 13-Apr-2020].

[9] A. Maqbool, “Coronavirus: why has the virus hit African Americans so hard?,” *BBC News*, 2020. [Online]. Available:

https://www.bbc.com/news/world-us-canada-52245690. [Accessed: 13-Apr-2020].

[10] J. Guy, "East Asian student assaulted in 'racist' coronavirus attack in London," *CNN.com*, 2020. [Online]. Available: https://www.cnn.com/2020/03/03/uk/coronavirus-assault-student-london-scli-intl-gbr/index.html. [Accessed: 13-Apr-2020].

[11] H. Yan, N. Chen, and D. Naresh, "What's spreading faster than coronavirus in the US? Racist assaults and ignorant attacks against Asians," *CNN.com*, 2020. [Online]. Available: https://www.cnn.com/2020/02/20/us/coronavirus-racist-attacks-against-asian-americans/index.html. [Accessed: 13-Apr-2020].

[12] M. Rajagopalan, "Korean interpreter says men yelling 'Chinese' tried to punch her off her bike," *BuzzFeed News*, 2020. [Online]. Available: https://www.buzzfeednews.com/article/meghara/coronavirus-racism-europe-covid-19. [Accessed: 13-Apr-2020].

[13] K. Schaeffer, "Nearly three-in-ten Americans believe COVID-19 was made in a lab," *Pew Fact Tank*, 2020. [Online]. Available: https://www.pewresearch.org/fact-tank/2020/04/08/nearly-three-in-ten-americans-believe-covid-19-was-made-in-a-lab/. [Accessed: 10-Apr-2020].

[14] R. Iyengar, "The coronavirus is stretching Facebook to its limits," *CNN Business*, 2020. [Online]. Available: https://www.cnn.com/2020/03/18/tech/zuckerberg-facebook-coronavirus-response/index.html. [Accessed: 13-Apr-2020].

[15] S. Frenkel, D. Alba, and R. Zhong, "Surge of virus misinformation stumps Facebook and Twitter," 2020. [Online]. Available: https://www.nytimes.com/2020/03/08/technology/coronavirus-misinformation-social-media.html. [Accessed: 13-Apr-2020].

[16] D. A. Broniatowski *et al.*, "Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate," *Am. J. Public Health*, pp. e1–e7, 2018, doi: 10.2105/AJPH.2018.304567.

[17] Y. Lama, T. Chen, M. Dredze, A. Jamison, S. C. Quinn, and D. A. Broniatowski, "Discordance between human papillomavirus Twitter images and disparities in human papillomavirus risk and disease in the United States: Mixed-Methods Analysis," *J. Med. Internet Res.*, vol. 20, no. 9, Sep. 2018, doi: 10.2196/10244.

[18] 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 *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, 2016, vol. 27, pp. 1361–1373, doi: 10.1145/2818048.2819927.

[19] N. F. Johnson *et al.*, "Hidden resilience and adaptive dynamics of the global online hate ecology," *Nature*, vol. 573, no. 7773, pp. 261–265, Sep. 2019, doi: 10.1038/s41586-019-1494-7.

[20] N. F. Johnson *et al.*, "New online ecology of adversarial aggregates: ISIS and beyond," *Science*, vol. 352, no. 6292, pp. 1459–1463, Jun. 2016, doi: 10.1126/science.aaf0675.

[21] Facebook, "Pages, groups and events policies," *Facebook Policies*, 2020. [Online]. Available: https://www.facebook.com/policies/pages_groups_events. [Accessed: 13-Apr-2020].

[22] S. Syed and M. Spruit, "Full-text or abstract? Examining topic coherence scores using Latent Dirichlet Allocation," in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2017, pp. 165–174, doi: 10.1109/DSAA.2017.61.

[23] "Latent Dirichlet allocation," *Wikipedia [English]*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation. [Accessed: 13-Apr-2020].

[24] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, 2015, pp. 399–408, doi: 10.1145/2684822.2685324.

[25] CDC, "CDC, Washington State report first COVID-19 Death," *Centers for Disease Control and Prevention*, 2020. [Online]. Available: https://www.cdc.gov/media/releases/2020/s0229-COVID-19-first-death.html. [Accessed: 01-Mar-2020].

[26] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014, pp. 63–70, doi: 10.3115/v1/W14-3110.

[27] P. D. Manrique, M. Zheng, Z. Cao, E. M. Restrepo, and N. F. Johnson, "Generalized gelation theory describes onset of online extremist support," *Phys. Rev. Lett.*, vol. 121, no. 4, p. 048301, Jul. 2018, doi: 10.1103/PhysRevLett.121.048301.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993967, IEEE Access

Richard Sear is pursuing the B.Sc. degree at the George Washington University with a major in Computer Science and minors in Mathematics and Physics. He is interested in unsupervised machine learning and natural language processing. In addition to his research at GW, 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.

Nicolás Velásquez received the B.A. in Political Science from Universidad de Los Andes (Colombia) in 2005, and the Ph.D. in International Studies from University of Miami in 2018. He is currently a Lecturer at the Elliot School of International Affairs and the inaugural post-doctoral Knight Fellow at the Institute for Data, Democracy and Politics, at the George Washington University in Washington D.C. He has been a fellow at the University of Miami's Center for Computational Sciences, and was an Andrés Bello scholar. His research interests include the application of computational social sciences and network sciences for conflict and public policy research. He is a founding partner of ClustrX LLC, and the Chief Data Scientist at Linterna Verde, a Colombian digital-literacy NGO.

Rhys Leahy received the B.A. from American University in International Studies. She is a research affiliate at the George Washington University's Institute for Data, Democracy and Politics, and a founding partner of ClustrX LLC. She was awarded scholarships from the U.S. Department of State to pursue advanced language studies in Russia and Tajikistan. She also previously worked as a science writer at the American Institute of Physics.

Nicholas Johnson Restrepo received the B.Sc. from Carnegie Mellon University in Business Administration in 2018. He is a research affiliate at the George Washington University's Institute for Data, Democracy, and Politics, and a founding partner of ClustrX LLC. He has business experience in small and large companies in Europe and South America.

Sara El Oud received the B.Sc. in physics from The American University of Beirut, Lebanon, and the M.Sc. in Physics from the George Washington University in 2020. She is currently a Ph.D. student in Physics at the George Washington University under the supervision of N.F. Johnson. Her research interests lie in the theory of complex systems and networks, in particular the role that heterogeneity and out-of-equilibrium dynamics play in aggregation processes in life science systems.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2993967, IEEE Access

Nicholas Gabriel received the B.Sc. in physics and mathematics from the University of Mary Washington. He is currently a Ph.D. student in Physics at the George Washington University under the supervision of N.F. Johnson. He is interested in connections between machine learning, network science, and statistical physics. Previously he has interned at Brookhaven National Laboratory and Massachusetts General Hospital.

Yonatan Lupu received the B.A. and J.D. from Georgetown University, and M.A. and Ph.D. from University of California-San Diego. He is a professor in the Department of Political Science at George Washington University. His current research projects focus on political violence, digital authoritarianism, human rights abuses, and violent online extremism.

Neil Johnson received the B.A. and M.A. from Cambridge University, U.K., and a PhD from Harvard University as a Kennedy Scholar. He is a professor in the Physics Department at the George Washington University. Before that, he was a Research Fellow at the University of Cambridge, and then a Professor of Physics at the University of Oxford until 2007, having joined the faculty in 1992. He presented the Royal Institution Lectures "Arrows of Time" on BBC TV in 1999. He is a Fellow of the American Physical Society and is the recipient of the 2018 Burton Award from the APS. His research interests lie in complex systems and networks.