Fighting Zika with Honey: An Analysis of YouTube's Video Recommendations on Brazilian YouTube

Jonas Kaiser, Adrian Rauchfleisch, and Yasodara Córdova

This paper analyzes the role of health misinformation around Zika on Brazilian YouTube. There is a trove of research on misinformation, disinformation, or "fake news" in the political—often Western European and U.S. American—context (Tucker et al., 2018). There is, however, less research on the role misinformation plays in other non-primarily political fields like health communication; especially in a country from the global south like Brazil. Furthermore, most research on misinformation focuses either on social media platforms like Twitter or Facebook (e.g., Guess, Nagler, & Tucker, 2019) or on exposure and/or sharing of news (e.g., Guess, Nyhan & Reifler, 2018). YouTube, in comparison, has been mostly ignored for research, although research is slowly catching up. The main focus of our case study is the "Zika" virus, i.e. the virus that is spread through mosquito bites and which caused a health crisis in Brazil in 2016 but is still around in 2019 (Jacobs, 2019). As 42% of Brazilians (Newman et al., 2019) get news from the Alphabet owned video platform, it is imperative to understand what type of content a simple search for "Zika" will display on YouTube and how the platform's recommendation algorithms contribute to the spread of health misinformation. In our analysis of YouTube's video recommendation algorithm, we show that there is a difference between the most prominent videos on Zika and the long tail of recommended videos: while the former consists out of mostly legitimate channels and information about Zika, the latter contains many videos from channels that promote conspiracy theories or alternative healing methods.

YouTube

Over 1.5 billion people use YouTube in a month (YouTube, 2017) and, consequently, it is an important cornerstone in the networked public sphere (Benkler, 2006). As stated above, 42% of Brazilians are using the video platform for news (Newman et al., 2019). However, YouTube is not only used for news but also for all sorts of entertainment (e.g., gaming, music, sports, etc.) and information (e.g., how to build computers or to do makeup, but also health information. Indeed, in an analysis on the most popular videos on Zika, Nerghes, Kerkhof, and Hellsten (2018) showed that about one third of the videos contained misinformation and that there was no significant difference in how users interacted with legitimate videos and those that contained misinformation. In a similar analysis, Bora et al. (2018) found that while there were more legitimate videos (~70%) on Zika in the top videos, that the videos containing misinformation had more views, likes, and shares. We add to this literature by focusing specifically on Brazil and on the networks that the recommendation algorithm creates.

Method

For our analysis, we collected the top 450 search results when searching for "Zika" on Brazilian YouTube. For our search we created a scraper on a Brazilian server and set Protuguese as interface language and Brazil as location in a non-personalized headless browser. We, then, collected the top 10 videos that YouTube recommended alongside these videos with YouTube's API. We chose the top 10 to simulate user behavior and while we could have collected even more recommended videos, we find it unlikely that users will scroll to the bottom of a video’s page to see all recommendations. In that line we also created channel networks for 1, 3, 6, and 10 recommendations. We then repeated this step once more for the videos that were added. To not only be able to understand how videos were related to each other through YouTube's recommendation algorithm but also the underlying structure, we aggregated the videos to their respective channels. In doing so, we both have networks for videos as well as channels where the connecting edges represent YouTube's video recommendation algorithm (see Figure 1).

In a next step we took all video descriptions, including video tags, and video titles and used this as the basis for a topic modeling analysis with the R package “stm” (Structural Topic Model; Roberts, Stewarts, & Tingley, 2017). We used stopword lists for Portuguese as well as English, removed all punctuation and numbers, and eventually stemmed the words in English, Spanish and Portuguese. Furthermore, we only used words that at least appeared 10 times or more in the whole corpus. While we tested the topic with 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, and 200 topics, we finally decided for 200 topics as this gave us the most granular number of relevant health related topics (n=13). We also calculated the overlap between videos based on their tags. Tags are chosen by the channel owners and usually are not visible for a user watching a video. We find that 46.5% of the video recommendations (directed edges) have for the source and the recommended video an overlap based on at least one tag (e.g., both videos have Zika as a tag), 31.8% of all edges have an overlap on two tags, 23.3% on three tags, 13.0% on at least 5 tags, and 6.3% on at least 10 tags (Fig. 2).

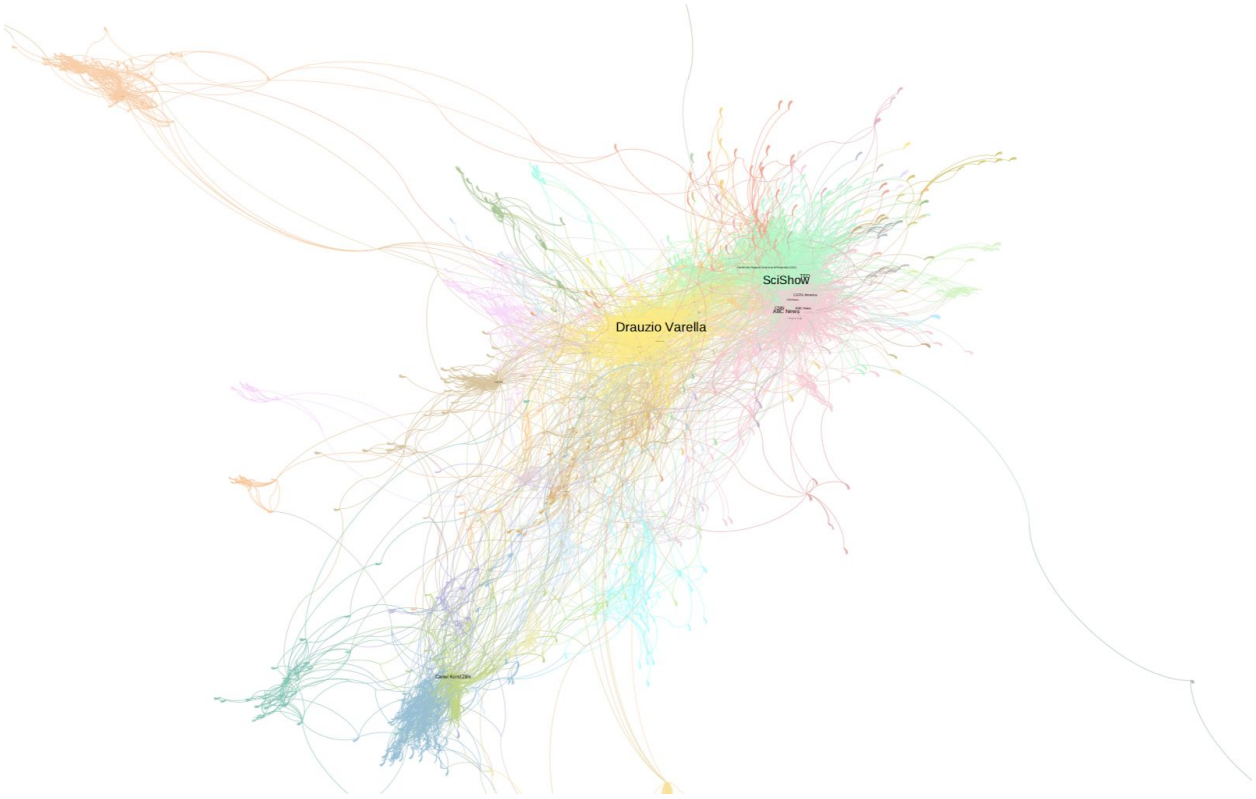


Figure 1. Channel network of Zika video recommendations (nodes=9,996, edges=19,512; node size per indegree).

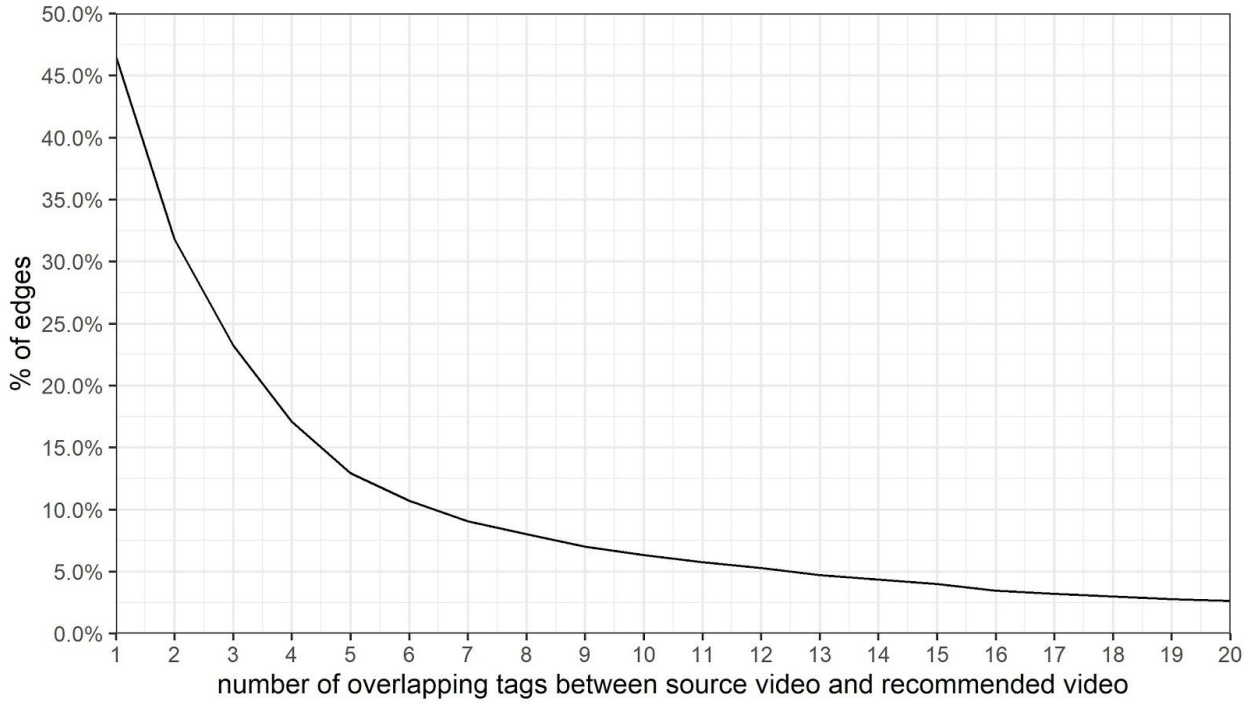


Figure 2. Overlap between videos based on N amount of tags.

Results

To make sense of the data we collected, we will present our findings in several steps. We will first quickly summarize the channel network which is based on the video recommendations. Then, we will present the video network. Finally, we will present the topic model and the associated topic model network.

Channels

We show that on Brazilian YouTube, the Zika videos can be roughly differentiate in three communities (Fig. 1): Brazilian videos on Zika (the yellow community), English-language videos on Zika (green and pink communities), and videos on the music artist Zika (dark green and blue communities). This already highlights YouTube's inherent international character as well as the diversity of content. However, when

analyzing the channel networks (Figure 1), we can differentiate between the most recommended channels (i.e. the channels which videos got recommended the most) and the so-called “long tail”, i.e. the channels that also got recommended but which received less recommendations. In our analysis of the 20 most recommended videos in our network, we were able to identify only one video that clearly contained misinformation. So while the most prominent channels contained mostly correct information on Zika, several of the smaller channels revolved around health misinformation. One video, for example, proposed several house remedies like honey to fight Zika. Indeed, perhaps the biggest finding in this content is that there is no clear delineation between channels that publish information on Zika and those that publish misinformation on a community detection level.

Videos

When not looking on the aggregate channel level but rather at the underlying video level, we can see that while the core recommendations of the videos are on the Zika virus, the recommendation algorithms quickly will also recommend unrelated content (Fig. 3). There are, for example, several music communities in the network which are mostly getting recommended through videos from the artist Zika. Similarly, there is a notable Serbian as well as a Yugoslavian TV community in the network. The algorithm’s logic, in short, connects videos on Zika in English to an old Yugoslavian TV show Žikina dinastija (or: Žika's Dynasty) and from there to Serbian politicas.

When focussing on the core network, i.e. the Brazilian and English-language communities, we see that there are several English-language communities, but only two core Brazilian communities (Fig. 4). The pink community on the right of Figure 4 is speaking about cures and remedies (e.g., Vitamin C) while the green community next to it is mostly focussing on Zika and other mosquito-transmitted diseases like Dengue or Chikungunya. Several of the less frequently recommended videos in both communities spread

misinformation: some spread myths about where Zika comes from while others will present home remedies against Zika in the form of garlic or honey. The community, then, that connects the Brazilian with the English-language communities consists of videos that talk about Zika in general "Zika Virus: What We Know (And What We Don't)" from the channel SciShow but also how mothers and families dealing with children who have Microcephaly which has been linked to Zika. From this community, YouTube recommended videos on Mosquitos as well as more general virus infections like Ebola. The blue community on the left, then, mostly consists of Tedx videos.

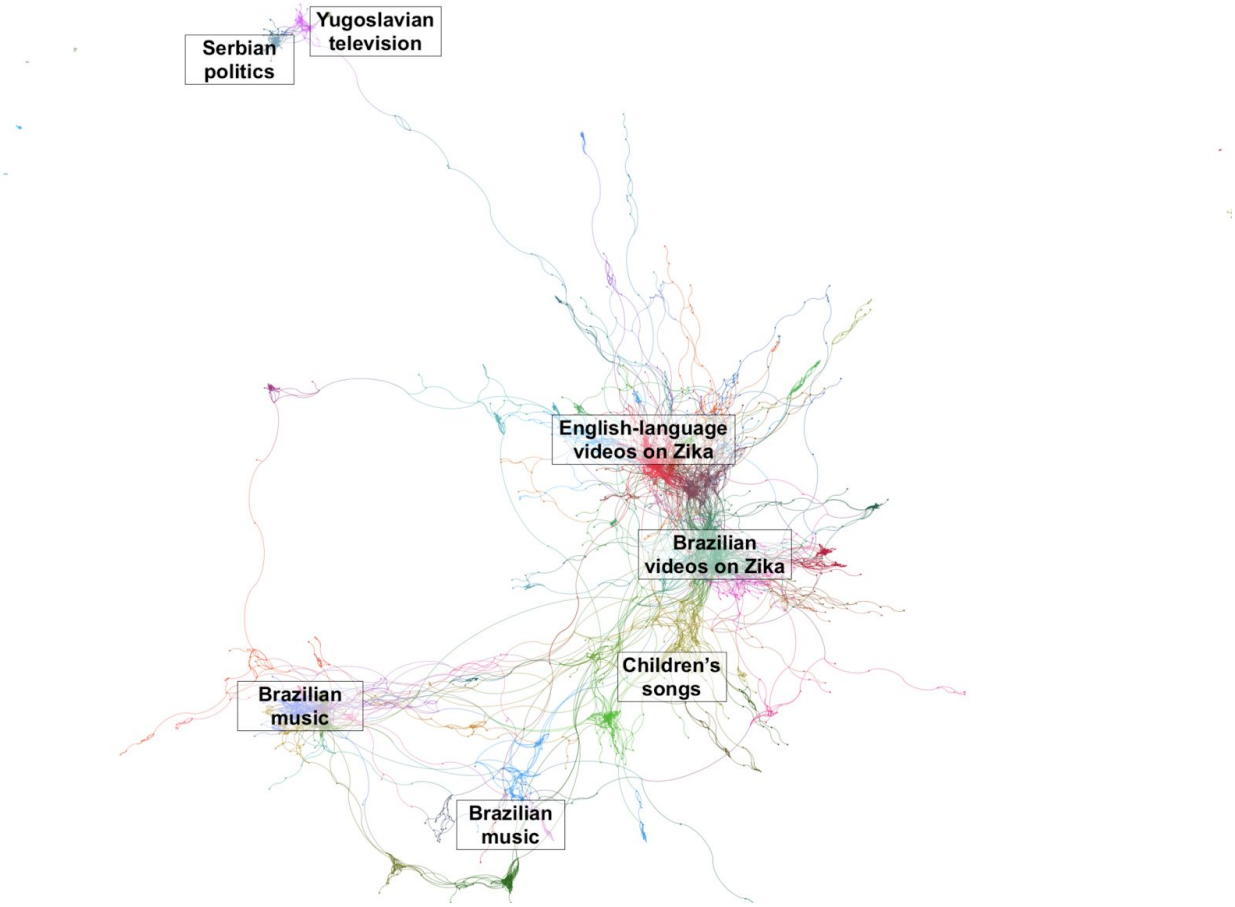


Figure 3. Video recommendation network (nodes = 20,242; edges = 26,091; community detection with Louvaine).

Figure 4. Core of video recommendation network (nodes = 3,978; edges = 5,882; community detection with Louvaine).

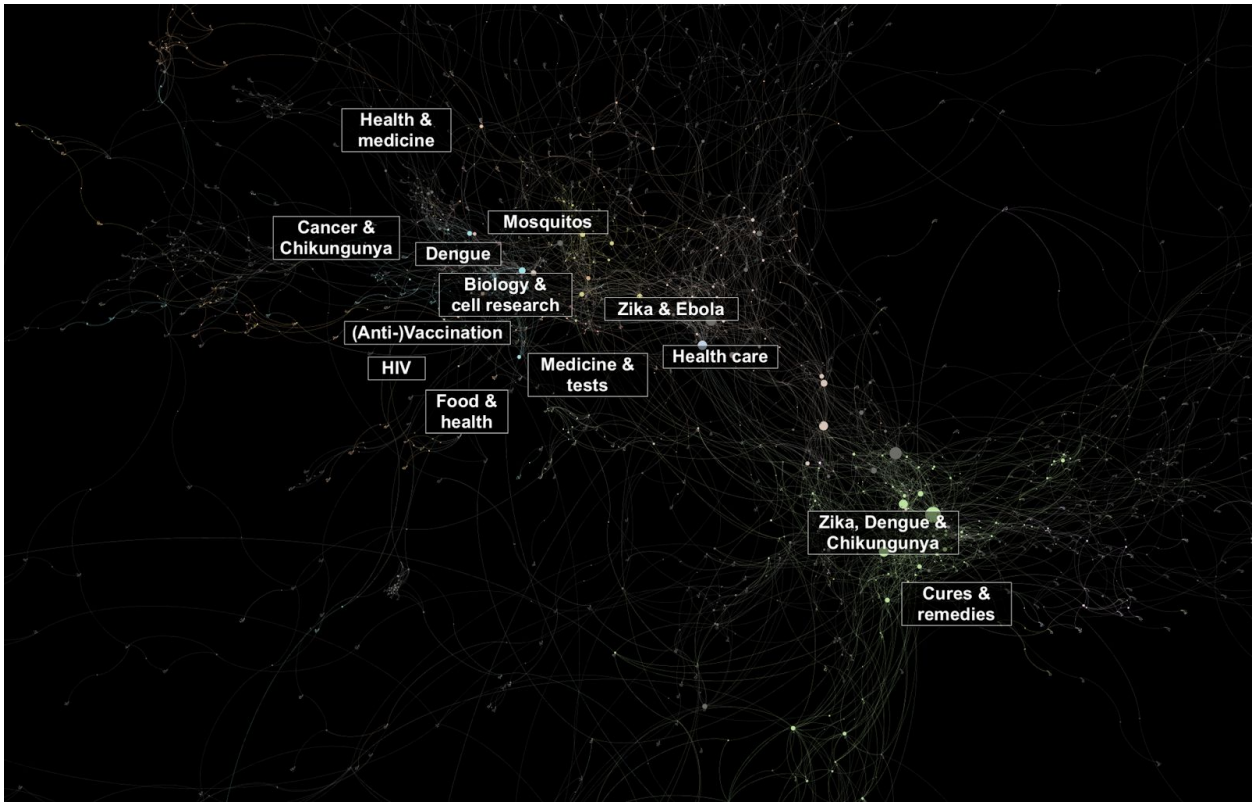
Topic modeling

To better understand what topics YouTube's algorithms were recommending we conducted a topic modeling analysis based on video descriptions, video tags, and video titles. We picked 200 topics to be able to identify more granular topics. After carefully looking through all 200 topics, we selected 13 topics that were the most relevant to our case (Table 1). In doing so we are able to highlight several key aspects: Zika is rarely discussed on its own on YouTube. We see Zika is being used in the same context as Dengue and Chikungunya but also Ebola. Furthermore, other disease topics include cancer (via Microcephaly) and HIV. In addition, the search for Zika also presented videos that talked about food, medicines, biology & cell research, health care, but also (anti-)vaccination standpoints (while some of the videos mock anti-vaccination talking points, others are titled like "New proof vaccines for Pharma profit, not health" from RT America).

Topic	Words with highest probability
Food & health	food, health, eat, diet, nutrit, cook, cholesterol
Health & medicine	dis, hospit, doctor, syndrom, surg, patient, infect
Biology & cell research	cell, research, biolog, scienc, immun, dna, institut
Dengue	blood, dengu, fev, pressur, sleep, health, high
Cancer & Chikungunya	canc, brain, treatment, tumor, health, rem, chikunguny
Cures & remedies	cas, remedi, agu, dor, natur, cur, receit
Health care	medic, prov, health, medicin, dis, caus, condit
Zika & Ebola	virus, zik, dis, health, infect, ebol, microcephal
Medicine & tests	test, elis, antibod, assay, posit, result, detect
Zika, Dengue & Chikungunya	dengu, zik, mosquit, virus, aed, chikunguny, doenc
Mosquitos	mosquit, mal, bit, repel, insect, control, larv
(Anti-)Vaccination	vaccin, anti, leav, health, asthma, autism, taw
HIV	hiv, aid, symptom, sign, treatment, dis, patient

Table 1. Relevant topics and words with highest probability.

To understand how the topics that we identified map on the video recommendation network, we identified all videos in the topic model where one of the 13 topics was the most prominent topic, i.e. it can be assumed that the video is most likely about that topic. We then imported the topics from the topic model into Gephi and colored the nodes based on the different topics (Fig. 5). In doing so, we can see that the topics do not only match the identified communities in Figure 4 but also give more context to what is being discussed in the core network. More specifically, we can see that, similar to the community detection, the Brazilian videos can be mostly distinguished between videos that talk about Zika and associated diseases on the one hand and cures and remedies on the other. The English-language videos, however, are much more fragmented and lead from one topic to another. But most importantly, here, too, we were not able to identify a misinformation topic. Indeed, while the long tail of the relevant videos seems to be full with misinforming videos, these are always connected to the overarching topic communities and, thus, trustworthy videos. Misinformation, in this sense, is always only one click away.



Conclusion

In this case study, we were able to show that health misinformation around Zika is a big issue on YouTube. And while the top video recommendations are mostly trustworthy, there is not only a lot of misinforming content on YouTube, it is also often only one click away. Indeed, we show in this case study that misinformation around Zika is not isolated to a potential filter bubble full of conspiracy theories but rather seems to reside in the recommendation algorithms long tail. This means that there is a lot of false content on YouTube and that this content will, eventually, getting recommended.

With that said, it is important to note that this analysis is based on one keyword and video recommendations. And while we tried to avoid personalization at all costs (tested both API as well as via browser, public and incognito mode, with VPN and without, etc.), it is likely that there is some bias in our

analysis. We are also intrigued by the finding that there are numerous health-related topics in the English-language topic model but only two topics for the Brazilian videos. While it is possible, that the topic model software that we used might work better for English, it is unlikely as we have used stm successfully for different non-English languages before. Another explanation might be that doing our research from a Brazilian server might lead to more fragmented English-language results (potentially due to prominence) which, while connected, are still distinct from each other. This explanation is also supported by the search results. In the search results at the beginning of our analysis only 3 videos in the top 20 and 25 videos in the top 100 are in English. However, starting from rank 101 up to 450 English videos were slightly more prominent than Portuguese videos. The lower the rank of a video in the search results, the more divergent is the content linguistically from the interface and location settings as well as topically from the intended interpretation of the search term.. This, in the end, also highlights the double edged sword that is the recommendation algorithm: often times contentious issues will be targeted by the creators of misinformation. And if there is little content, YouTube will recommend whatever is available, no matter the source. This is also what Golebiewski and boyd (2018) call "data voids." The issue of recommendation algorithms goes beyond that, however, since on YouTube there is no "end" to the recommendations and, thus, bad content will eventually get recommended. We see this in our study where the most prominent search results and recommendations are mostly legitimate, the problem, however, is the long tail.

Indeed, our research highlights the importance of conducting research on YouTube. Not only to understand what is happening but also to force YouTube to be more transparent about their algorithms and how they influence how people watch news.

Literature

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven: Yale University Press.

Bora, K., Das, D., Barman, B., & Borah, P. (2018). Are internet videos useful sources of information during global public health emergencies? A case study of YouTube videos during the 2015–16 Zika virus pandemic. *Pathogens and Global Health*, 112(6), 320–328. https://doi.org/10.1080/20477724.2018.1507784

Golebiewski, M., & boyd, danah. (2018). Data Voids: Where Missing Data Can Easily Be Exploited (p. 7). Retrieved from Data&Society website: https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf

Guess, A. M., Nyhan, B., & Reifler, J. (2018). *Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign*. European Research Council. Retrieved from http://www.ask-force.org/web/Fundamentalists/Guess-Selective-Exposure-to-Misinformation-Evidence-Presidential-Campaign-2018.pdf

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1). doi:10.1126/sciadv.aau4586

Jacobs, A. (2019, July 2). The Zika Virus Is Still a Threat. Here's What the Experts Know. *The New York Times*. Retrieved from https://www.nytimes.com/2019/07/02/health/zika-virus.html

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2019). Reuters Institute Digital News Report 2019. *Reuters Institute Digital News Report*, 156.

Nerghes, A., Kerkhof, P., & Hellsten, I. (2018). Early Public Responses to the Zika-Virus on YouTube. In Proceedings of the *10th ACM Conference on Web Science - WebSci '18*. ACM Press. https://doi.org/10.1145/3201064.3201086

Roberts, M. E., Stewart, B. M., & Tingley, D. (2017). stm: R Package for Structural Topic Models. *Journal of Statistical Software*, 42.

Tucker, J., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., … Nyhan, B. (2018). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3144139

YouTube. (2017). *Updates from VidCon: more users, more products, more shows and much more*. Retrieved from https://youtube.googleblog.com/2017/06/updates-from-vidcon-more-users-more.html

Biography

Jonas Kaiser is an Affiliate at the Berkman Klein Center for Internet & Society and Associate Researcher at the Humboldt Institute for Internet and Society in Berlin. At Berkman Klein, he's heading the misinformation working group. His research is focused on political online communication from a comparative perspective.

Adrian Rauchfleisch is an Assistant Professor at the Graduate Institute of Journalism, National Taiwan University. His research is mainly focused on political online communication.

Yasodara Córdova is a Fellow at the Digital Kennedy School and Affiliate at the Berkman Klein Center for Internet & Society. She is also a fellow at the Center for Technology and Society (CTS) at FGV School of Law. At DKS, she is doing research on Data for public Policies and Digital Identity.