YTTREX: crowdsourced analysis of YouTube's recommender system during COVID-19 pandemic

Leonardo Sanna[1], Salvatore Romano[2], Giulia Corona[3], Claudio Agosti[4]

[1] University of Modena and Reggio Emilia, Modena, Italy
leonardo.sanna@unimore.it
[2] University of Padova, Padua, Italy
salvatore.romano.3@studenti.unipd.it
[3] University of Milano, Milan, Italy
giulia.corona1@studenti.unimi.it
[4] Tracking Exposed, https://tracking.exposed/
claudio@tracking.exposed

Abstract. Algorithmic personalization is difficult to approach because it entails studying many different user experiences, with a lot of variables outside of our control. Two common biases are frequent in experiments: relying on corporate service API and using synthetic profiles with small regards of regional and individualized profiling and personalization. In this work, we present the result of the first crowdsourced data collections of YouTube's recommended videos via YouTube Tracking Exposed (YTTREX). Our tool collects evidence of algorithmic personalization via an HTML parser, anonymizing the users. In our experiment we used a BBC video about COVID-19, taking into account 5 regional BBC channels in 5 different languages and we saved the recommended videos that were shown during each session. Each user watched the first five second of the videos, while the extension captured the recommended videos. We took into account the top-20 recommended videos for each completed session, looking for evidence of algorithmic personalization. Our results showed that the vast majority of videos were recommended only once in our experiment. Moreover, we collected evidence that there is a significant difference between the videos we could retrieve using the official API and what we collected with our extension. These findings show that filter bubbles exist and that they need to be investigated with a crowdsourced approach.

Keywords: Algorithm analysis, crowdsourced data collections, network analysis, official API, COVID-19, YouTube, filter bubble.

1 Introduction

1.1 Algorithmic personalization

Algorithmic personalization is now part of our lives. In fact, recommendation systems are used for a remarkably high number of tasks, ranging from working to free time.

Each time we google something, an algorithm is selecting what is most relevant for us, the same happens when we scroll our Facebook feed and when we use our Netflix or Spotify account. We may say that algorithms are the technological solution to the information overload we live on daily.

However, most of these services are owned by private corporations that use black-box algorithms to curate the content selection for their users. In the last few years, these platforms have been at the stake of academic research, particularly for what concerns the so-called "fake news debacle", with a lot of research focusing on misinformation spreading and on the polarization of the public debate[1, 2, 3].

For the sake of clarity, one crucial distinction has to be made about this. Research that considers information spreading, online debate and user engagement is a study on *echo chambers*. Although not well defined in the literature, echo chambers are social phenomena based on ideological affinity and are relatively accessible for research. Instead, in this paper, we are interested in studying the *filter bubble* [4], which is the direct effect of algorithmic personalization. To use Zimmer's words [5], echo chambers are created by users, while filter bubbles are made by algorithms.

After the concept became widely discussed in the everyday debate, the academics started questioning the existence of the filter bubble effect, following the ideas proposed by Bruns [6] that claimed that there is no evidence that echo chambers and filter bubbles exist outside the academic theory. Furthermore, studies on filter bubbles and echo chambers focused almost exclusively on the polarization of the debate, overlooking other more critical areas of inquiry such as the lack of tools to account for user personalization. Empirical research on algorithmic personalization is still quite fragmented and we believe that this happens because of the lack of a shared methodology among the researches. The fact that we must take into account user experience is problematic because of the number of uncontrollable variables and also because it requires user collaboration or fabricated profiles.

For its part, YouTube provides information about its algorithm, but just related to the general structure of the recommended system algorithm [7, 8]; thus, we cannot state for sure how many variables are used for the personalization process and how. The site also provides an official API, often used by researchers, but this does not include individual personalization, and, as we will show in section 4.2, API data might differ a lot from the actual users' experience.

In the following section, we review the latest methods for algorithmic personalization analysis, then we illustrate our methodology and the YTTREX tool. Finally, we discuss the results of our experiment and the main limitations of our work.

1.2 Methods for algorithmic personalization analysis.

There are a number of studies that claim to be focused on filter bubbles. However, the majority of these works is not actually taking into account algorithmic personalization but, instead, it is inquiring ideological preference (echo chambers) with social sciences methods [9, 10].

These approaches do not consider the fact that algorithmic personalization is a product of the interaction between users and platforms and it is essentially passive

since users have extremely limited or no control of their personalization. Hence, if we approach filter bubbles starting from user behavior, we are not studying filter bubbles but echo chambers. In our opinion, trying to infer conclusions on algorithmic personalization starting from its alleged effects might produce misleading outcomes.

In this work we are proposing a crowdsourced approach, as it was already experimented by Robertson et al. on Google SERP [11]. In their study the authors stated that they found truly little evidence of filter bubble effect. Although we found their methodology robust and well-suited for the study of filter bubbles, we believe that their conclusions should be reconsidered taking into account that:

1. algorithmic personalization is platform-specific, as every platform has its own algorithm
2. Google SERP might not be the best platform to inquiry filter bubble in its original meaning of "informational bubble", as it is produced by an intentional research
3. The "in incognito mode" of the web browser Chrome is not a completely clean navigation, as it keeps a certain level of personalization that is based, for instance, on geographic location.

Regarding point (1), during one of our previous experiments we found evidence of algorithmic personalization on Facebook using synthetic profiles [12]. We used this approach so that we could control all the variables. Regarding point (3) we must recall that controlling the variable is one of the main difficulties while studying algorithmic personalization, also because we have to proceed by trials and errors, since we do not know exactly which variables influence algorithmic decisions.

Hence, we propose to investigate specific effects (such as ideological preference) using fabricated profiles and use crowdsourced approach to gain evidence of existence of personalization, namely to account for the fragmentation of content distribution.

Measuring filter bubbles on YouTube. We propose a crowdsourced methodology to investigate and measure user personalization within the recommended videos on YouTube (YT). The context of COVID-19 was a unique occasion to test our tool YTTREX, looking at how YouTube distributed its content related to the pandemic. An extensive review on YouTube research has been conducted by Arthurs et al. [13], who highlighted that the platform is vastly understudied. Other studies on YT filter bubbles that propose a crowdsourced approach do not currently exist to the best of our knowledge. Related works use YT API or other methods that are not user-centered, nor they collect empirical data directly from users' browsers [14, 15, 16, 17]

2 Tool: YouTube Tracking Exposed

2.1 How it works

The browser extension (add-on) of Tracking Exposed[1] collects evidence from the metadata that is observable on the web page when the user lands on the homepage, watches a video, or does research on the YouTube website. It creates cryptographic key pairs to ensure the user can access her/his data. It is necessary because the tool does not have an email address, Google profile, or any other authentication method based on personal data. The tool collects separate contributions for each browser with the add-on installed.

The data are collected in three phases:

1. **Collection**: the add-on takes a copy of the HTML when the browser is watching a video. Four buttons appear on the top left of the screen (Fig.1), when the add-ons is installed and enabled by the popup. The color code represents the different status.

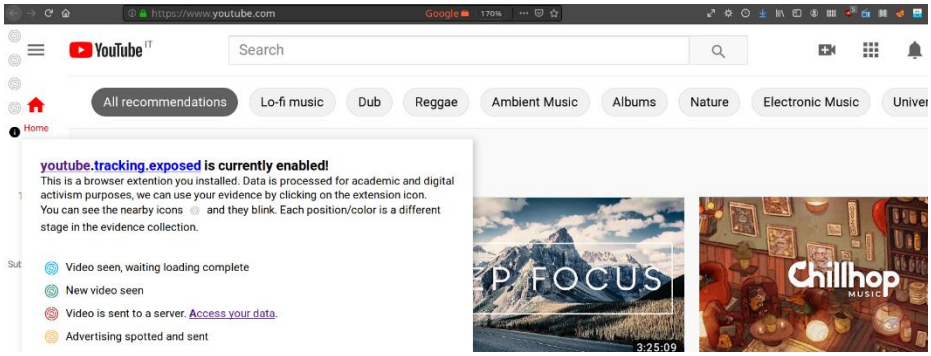

Fig. 1. Screenshot of what the browser extension shows while navigating on YouTube.

2. **Parsing**: server side, the HTML is processed and metadata are extracted. The information is then organized in a dataset. In the HTML there are many different data that might be analyzed to extract metadata. We did not yet extract all possible information, especially we avoided any unique tracker that might become personal data if collected. On the other hand, the YTTREX project still has room for improvement, and we might not have yet mapped 100% of the potentially interesting metadata for YouTube algorithm analysis.

1 https://youtube.tracking.exposed, AGPL3 code: https://github.com/tracking-exposed/yttrex/

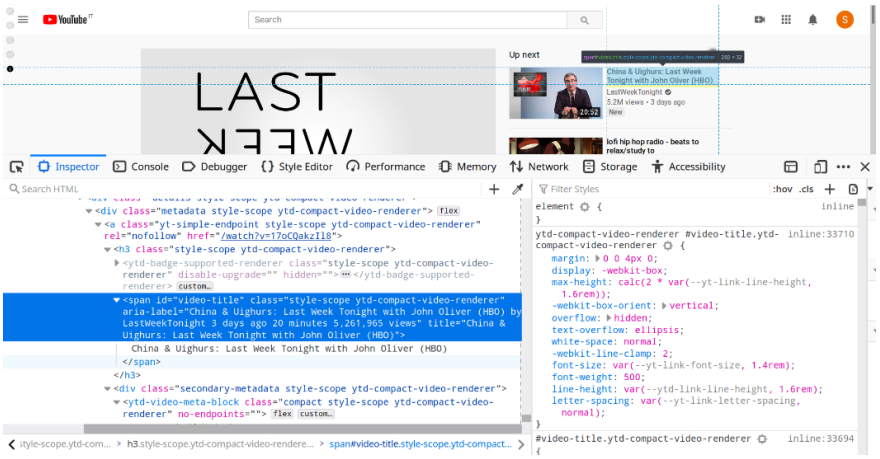

Fig. 2. HTML inspection of a recommended video on YouTube and its aria-label

We were also able to record the users' interface, detect the language, record related videos, the number of views, and duration. Inspecting the HTML of a recommended video (Fig. 2), you might see the data field named aria-label[2]. This text field is meant for accessibility and contains a compacted, but human formatted, set of information useful for researchers. Because of the localization, YouTube produces aria-label with strings that change accordingly to the user interface Language. For example, the aria-label: "*Crise pétrolière : coup de poker sur l'essence | ARTE by ARTE 6 days ago 58 minutes 213,982 views"* is composed by the information shown in Table 1.

Title	Crise pétrolière : coup de poker sur l'essence | ARTE
UX Language dependent *stopword*	by
Publisher name	ARTE
Relative human readable publication time	6 days ago
Human readable video length	58 minutes
Number of views formatted as per UX locale standard	213,982 views

Table 1. Aria-label composition.

We might externalize this natural language conversion, managed by our aria-label parsing library[3], as an independent library, once we figure out how to maintain the list of fixed terms that scale up proportionally to the language supported by YouTube. The sum of session information, video watched, and recommended videos, produces the data unit with the format detailed in Table 2.

[2] For reference see: https://mzl.la/33dMuRN

[3] https://github.com/tracking-exposed/yttrex/blob/master/backend/parsers/longlabel.js

Field Name	Data Type	Description
login	Boolean	True if the profile was logged on YT
id	String	Unique identifier for each installed extension
savingTime	ISODate	GMT hour when evidence get saved
clientTime	ISODate	Date on the users' browser
uxLang	ISO 639-1 code	Browser language
recommendedId	String	Unique identifier of the data unit
recommendedVideoId	String	Video unique ID used in YT URL
recommendedAuthor	String	Publisher of the recommended video
recommendedTitle	String	Title of the recommended video
recommendedPubTime	ISODate	Date of recommended video publication
recommendedRelativeS	Number	Seconds between recommended publication and access to watch the video
recommendedViews	Number	Views at *savingTime* for the recommended video
recommendedForYou	Boolean	True if YT explicitly says recommended for you
recommendedVerified	Boolean	True if publisher has the blue check ✔
recommendedKind	String	Live streaming or video
recommendedLength	Number	Duration of the video in seconds
recommendedDisplayL	String	Human formatted duration of video
watchedVideoId	String	From YT URL, the Video ID
watchedTitle	String	Title of the watched video
watchedAuthor	String	Publisher of the watched video
watchedChannel	String	Relative URL of YouTube channel
watchedPubTime	ISODate	Publication time of the watched video
watchedViews	Number	Amount of views at *sav-*

		ingTime
watchedLike	Number	Amount of thumbs up at *savingTime*
watchedDislike	Number	Amount of thumbs down at *savingTime*
sessionId	String	Unique identifier of users' sequence
hoursOffset	Number	Amount of hours after the 25 March 2020 GMT, the beginning weTest1
experiment	String	'weTest1', the experiment of this paper
pseudonym	String	A unique pseudonym for each browser plugin
top20	Boolean	True if *recommendationOrder* < 20
isAPItoo	Boolean	True if recommended is also in YT API related
step	String	Human readable language of watched video

Table 2. Data structure

3. **Research and data-sharing**: YTTREX was created to support independent analysis and privacy-preserving sharing of the algorithmically powered circulation of videos. Every video observation has a dynamic number of related videos (if the watcher scrolls the video page down, the browser loads 80 or more related videos, but for users who do not scroll down the default is to receive and display only the first 20 related videos). Every related video becomes a single row, a data record with its own unique ID. Interconnecting these with *metadataId*, the researcher might re-group all the related videos belonging to the same evidence, as they were displayed to the watcher. Certain fields such as logged, pseudo, and *savingTime*, are the same across the same id because they depend on the collection condition. *recommendedVideos*, *recommendedAuthor*, and other recommended-fields, changes in each row according to the related video described*; recommendedId* is generated for each row and should be used as guarantee of unique field.

According to the definition provided by Sandvig [18], the tool enables the user to potentially four of the five methods of algorithmic audit: Noninvasive User Audit, Scraping Audit, Sock Puppet Audit, if they have the know-how to use bots, and Crowdsourced or Collaborative Audit, as the experiment presented on this paper.

The database collected for this paper is available on Tracking Exposed website[4] and the code is available on GitHub[5] protected by AGPL v3 license.

[4] https://youtube.tracking.exposed/data/

[5] https://github.com/tracking-exposed/youtube.tracking.exposed

3 Experiment

A first test of our methodology was conducted during the Digital Methods Initiative Summer School in 2019[6]. For the first time, we compared the different users' experience in that experiment, manipulating some variables such as logged profile vs unlogged. This experience has been used as a baseline for the experiment design of this study.

3.1 Experiment design

We made a call for participation on our website to select the participants[7]. Every participant joined the experiment for free and voluntarily. The procedure involved the visualization of five videos about COVID-19 prevention, produced by the BBC channel, one for each of the most spoken languages in the world: Chinese, Spanish, English, Portuguese, Arabic. We picked these videos because we wanted to find a source equally trustworthy in the five languages. Originally the idea of this experiment comes from our doubt that YouTube could not effectively take down conspiracy theory on COVID-19[8], differently to what is claimed. We suspect English language and recommendation might benefit from a better curation, thus by comparing the recommended videos close to equally accurate COVID-19 videos. Still, in different languages, we could neither confirm nor reject the hypothesis.

We did not provide additional information about the minimum time that had to be spent watching the videos: loading the page was enough to collect the HTML. Participants could choose to perform the test logged with their personal account or without, the tool records if the user is logged or not, without collecting any data related to the specific account.

3.2 Official API comparison

The same day of the test, we retrieved via the official YouTube's API the related videos for the five videos included in the methodology.

Since language is an option for the API request, we performed five requests, one for each language. 50 videos were retrieved in each API request. We then stored this information using the metadata *isAPItoo* (see Table 2) for each of our evidence collected via YTTREX.

---

6 https://wiki.digitalmethods.net/Dmi/SummerSchool2019AlgorithmsExposed

7 https://youtube.tracking.exposed/wetest/1

8 https://www.nytimes.com/interactive/2020/03/02/technology/youtube-conspiracy-theory.html

4 Findings

4.1 Evidence of filter bubbles

The distribution of the recommended videos is clearly skewed as shown in Fig. 3. We investigated the distribution of recommended videos taking into account the language of the starting video, the browser's language, and considering whether the user was logged or not. No matter of which variable we took into account we always obtained a skewed distribution, as shown in the example of Fig. 4.

Fig. 3. Frequency distribution of recommended video in our dataset.

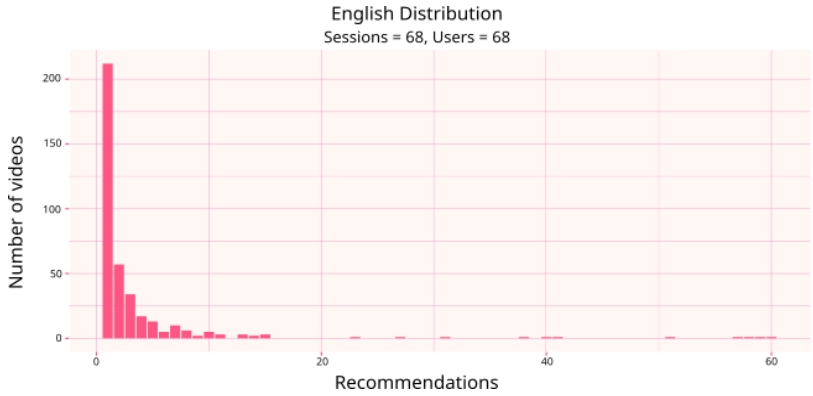

Fig. 4. Frequency distribution of recommended videos starting from the BBC video in English.

Our findings show that the vast majority of videos are recommended very few times (1-3 times), regardless of the variable considered. This distribution is significantly positively skewed according to Fisher’s skewness coefficient (>2). Summing up, 57% of the recommended videos have been recommended only once and only around 17% of the videos have been recommended more than 5 times during our experiment.

These results highlight that the filter bubble is real, and that algorithmic personalization produces a high fragmentation of recommended content among YouTube users. Another relevant finding for the study of algorithmic personalization is the huge difference we found using YT API and our tool. For users logged into their Google account only 11% of the recommended videos could be retrieved using the API as showed in the following section, in Fig. 8.

Finally, we calculated Lorenz curve over the distribution of the recommended video, confirming that the inequality in the distribution (Gini > 0.5) of the recommended videos.

We also calculated the Gini index for the number of videos selected for each user, since the result shown in Fig. 5 might be caused by an uneven number of videos selected for each user. However, with a Gini coefficient around 0.2, we have evidence that the algorithm is selecting an equal number of videos for each user, while distributing unevenly the recommendations for each video.

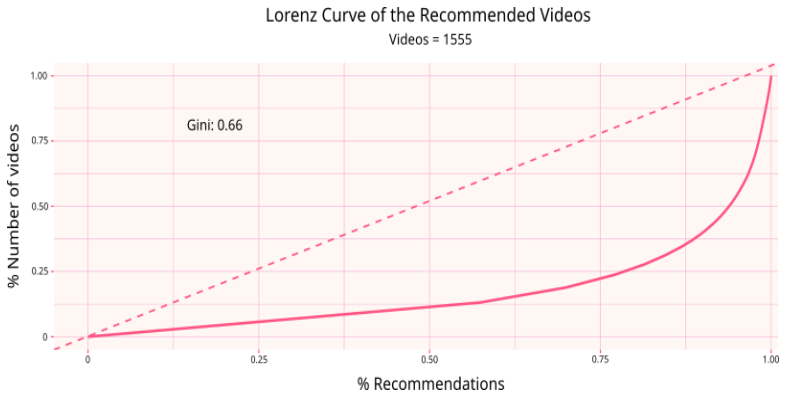

Fig. 5. Lorenz Curve and Gini Coefficient of the recommended videos

4.2 Network analysis

We performed a network analysis using Gephi [19] to better understand and visualize how the recommender system creates a filter bubble around users watching the same video the same day. Thanks to the Medialab's tool *Table2net*[9] we extracted a network file from the csv file. We created a bipartite network linking two types of nodes: users' pseudonyms and suggested video's ID.

In the graphs (Fig. 6, 7, 8) we used a circular layout algorithm [20] to dispose of all the users in a circle. We aimed to show all the participants in the same positions, pointing in the same direction, because they were performing the same task: in the examples they are watching the video from the English version of BBC channel "How do I know if I have coronavirus? - BBC News.". This representation allowed us to show how, even if they were all watching the same video, they were getting a different configuration of suggested videos.

9 https://medialab.github.io/table2net/

Then, we performed a Force Atlas 2 Algorithm to place each related video close to the users who received that suggestion by the platform. On the one hand this technique highlights the network centrality of several videos homogeneously suggested across users (the ones in the center of the graph); on the other hand we can clearly see videos suggested to singular users (those who are external to the users' circle, really close to just one pseudonym).

The size of the nodes is based on the degree of each node: in a range between size 15 and size 60, each user and each recommended video is big in relation to the number of links that it has. The videos in the center of the graph are bigger because they have been suggested more than the others. Because of a graphical compromise, the nodes with a degree minor than 15 have the same shape, likewise the nodes with a degree higher than 60 are all the same.

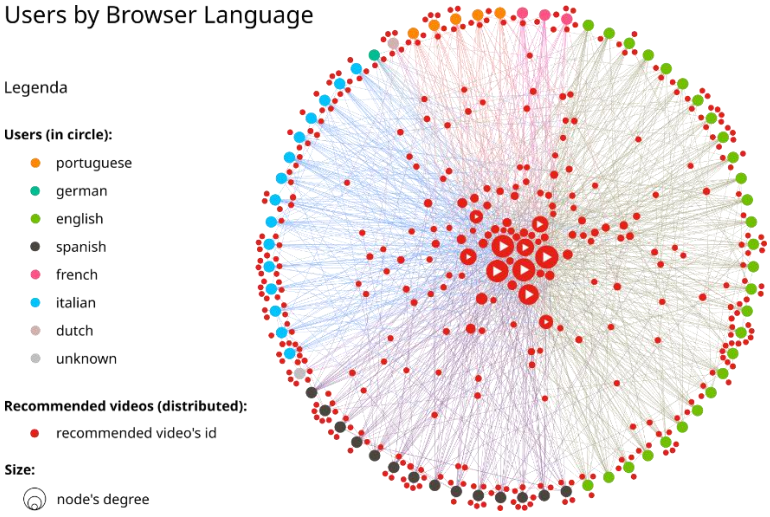

Fig. 6. Graph of the videos suggested to the participants while watching the video "How do I know if I have coronavirus? - BBC News**"**

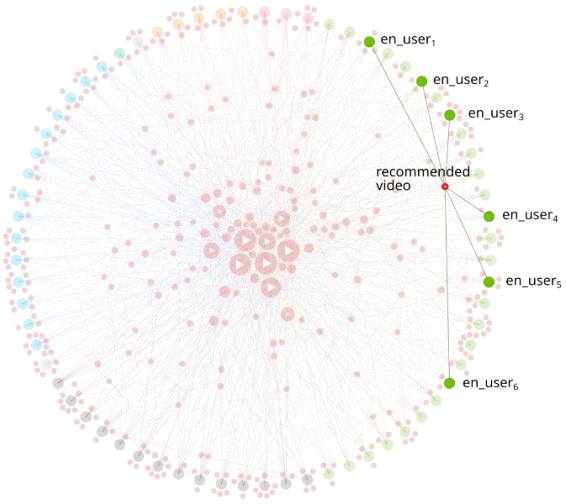

Fig.7. Zoom of Fig. 6, an example of video suggested only to users with English interface.

In Fig. 7 we highlighted how some of the videos recommended appear only to users with English browsers. This shows that the participants in the experiment received personalized suggestions according to their characteristics, despite watching the same video. This type of analysis can demonstrate differences in the users' experiences tracing the most influential features that can generate changes in the platform experiences.

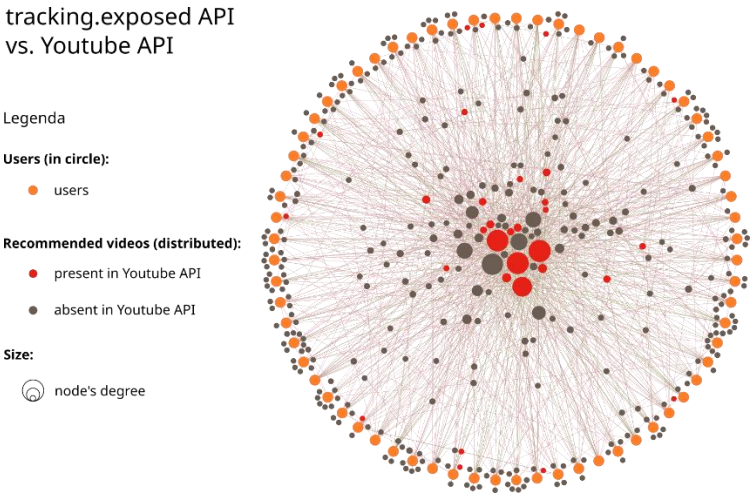

Fig.8. Same graph of Fig. 6, here the colors highlight the differences between the videos recorded with Tracking Exposed and the ones retrieved with YouTube official API.

As we already said in the previous section, there is a huge difference between the recommended videos that we retrieved from the API and the actual recommendations

(Fig. 8). The majority of the videos retrieved by the Tracking Exposed tool are not present in the database created with YouTube's API. Some of the most suggested videos (biggest nodes in the center of the graph) neither. This is relevant because it is evidence against the usability of official YT's data in academic research. The official API cannot represent the real variance of suggestions present in the actual recommended videos. Many scientific articles [21, 22, 23, 24] rely on these data to explain the circulation of videos on the platform, but according to our findings we might say that API data are just a generic representation of an ideal user that is really difficult to find in reality (no one of the users in our experiments gets the same recommendations as in the API).

The official API does not represent the various levels of personalization that occur in relation to the structural users' characteristics and to their past online behaviors. Thus, we cannot use API data to make inferences about personalization, polarization and filter bubbles, because these phenomena presuppose the study of real users in real context.

5 Conclusions

This research was intended as a proof of concept work, being the very first crowdsourced experiment carried out with our tool. It is not possible at this stage of our research to further generalize our findings, nor we can go in deep with content analysis, as our sample was quite small (68 users). Nonetheless, we gained evidence of the existence of algorithmic personalization on YouTube, what is also called filter bubble. Our dataset has been collected in a scenario in which we expected a shared common ground among users, since they all started from the same video on how to prevent COVID-19 infections. Instead, our data show a strong fragmentation of content selection, suggesting that there might be a lack of shared information on COVID-19 on YouTube.

We propose to measure the filter bubbles as we did in section 4.1; when the distribution of recommended videos is significantly positively skewed, we have evidence of filter bubbles. We also propose to measure the filter bubble using the Gini coefficient; a higher value of the Gini index would indicate a stronger filter bubble. Our analysis proves the necessity of further investigation on algorithmic personalization with crowdsourced and independent tools. In fact, our research also showed that official APIs cannot retrieve the majority of videos that are shown to the final users.

Further research on these themes should focus on 1) repeating the experiment with a larger sample 2) qualitatively explore the recommendations 3) study other structural characteristics of the users, better understanding the effects on the recommendations system, as we have already done with visualization time on similar recommender systems 4) investigate the differences in the home page and in the search engine of the platform, as recently done by others groups 5) analyse the levels of curation in different languages, comparing the percent numbers of fake news or conspiracy videos. The distributed approach success also depends on the number of volunteers participating

in the experiment. We consider this an issue of outreach, campaigning, and how the promoters frame the investigation to raise interest in key online communities.

References

1. Fernandez, M., Harith A:, Online Misinformation: Challenges and Future Directions. In Companion Proceedings of the The Web Conference 2018, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 595–602 (2018).
2. Zollo F., Bessi A., Del Vicario M., et al.: Debunking in a world of tribes. PLoS ONE 12(7): e0181821. (2017).
3. Del Vicario M., Vivaldo G., Bessi A., et. al.: Echo chambers: Emotional contagion and group polarization on Facebook. Scientific reports,6:37825 (2016).
4. Pariser, E.: The filter bubble: What the internet is hiding from you, Penguin UK.P, (2011).
5. Zimmer, F., Scheibe K., Stock M., et. al.: Fake news in social media: Bad algorithms or biased users?. Journal of Information Science Theory and Practice 7(2): 40-53 (2019).
6. Bruns, A.: Filter bubble. Internet Policy Review, 8(4) (2019).
7. Covington, Paul, Jay Adams, and Emre Sargin "Deep Neural Networks for YouTube Recommendations". Proceedings of the 10th ACM conference on recommender systems. ACM (2016).
8. Zhe, Z., Lichan, H., Li, W., Jilin, et al. : Recommending what video to watch next: a multitask ranking system. In Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19). Association for Computing Machinery, New York, NY, USA, 43–51. (2019)
9. Trielli D., Diakopoulos, N.: Partisan search behavior and Google results in the 2018 U.S. midterm elections. Information, Communication & Society DOI: 10.1080/1369118X.2020.1764605 (2020).
10. McKay, D., Makri, S., Guiterrez-Lopez, M, et al.: ACM, New York, NY, USA, We are the Change that we Seek: Information Interactions During a Change of Viewpoint. In Proceedings of ACM Conference on Human Information Interaction and Retrieval (CHIIR'20)., 10 pages. https://doi.org/10.1145/1234567890 (2019).
11. Robertson, R., E., Jiang, S, Joseph, K., et. al.: Auditing Partisan Audience Bias within Google Search. Proeedings of ACM Human.-Computer. Interaction 2, CSCW, Article 148 (November 2018), 22 pages. DOI: 10.1145/3274417 (2018).
12. Hargreaves, E., Agosti C., Menasché D., et al.: Biases in the Facebook News Feed: a Case Study on the Italian Elections. International Conference on Advances in Social Networks Analysis and Mining, August 2018, Barcelona, (2018)
13. Arthurs, J., Drakopoulou, S., & Gandini, A. : Researching YouTube. Convergence, 24(1), 3–15 (2018).
14. Song, M., Yun, J., Anatoliy, G..: Examining sentiments and popularity of pro-and anti-vaccination videos on YouTube. In Proceedings of the 8th International Conference on Social Media & Society, pp. 1-8. (2017).
15. Abisheva, Adiya, David Garcia, and Frank Schweitzer. "When the filter bubble bursts: collective evaluation dynamics in online communities." In Proceedings of the 8th ACM Conference on Web Science, pp. 307-308.(2016).
16. Bishop, S.: Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm. Convergence, 24(1), 69–84 (2018).

17. Rieder, B., Matamoros-Fernández, A., & Coromina, Ò.: From ranking algorithms to 'ranking cultures': Investigating the modulation of visibility in YouTube search results. Convergence, 24(1), 50–68. (2018).
18. Sandvig C, Hamilton K., Karahalios K., Langbort C.: Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. DATA AND DISCRIMINATION: CONVERTING CRITICAL CONCERNS INTO PRODUCTIVE INQUIRY, A PRECONFERENCE AT THE 64TH ANNUAL MEETING OF THE INTERNATIONAL COMMUNICATION ASSOCIATION. May 22, 2014; Seattle, WA, USA. (2014).
19. Bastian, M., Heymann, S., & Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In Third international AAAI conference on weblogs and social media. (2009).
20. Six, J. M., & Tollis, I. G.: A framework and algorithms for circular drawings of graphs. Journal of Discrete Algorithms, 4(1), 25-50. (2006).
21. Brbić, M., Rožić, E., & Žarko, I. P.: Recommendation of YouTube Videos. In 2012 Proceedings of the 35th International Convention MIPRO (pp. 1775-1779). IEEE. (2012).
22. Ledwich, M., & Zaitsev, A.: Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. arXiv preprint arXiv:1912.11211. (2019)
23. Marchal, N., Au, H., & Howard, P. N.: Coronavirus news and information on YouTube. Health, 1(1), 0-3. (2020)
24. Airoldi, M., Beraldo, D., & Gandini, A.: Follow the algorithm: An exploratory investigation of music on YouTube. Poetics, 57, 1-13 (2016):.