How YouTube Leads Privacy-Seeking Users Away from Reliable Information

Larissa Spinelli
Boston University
lspinell@bu.edu

Mark Crovella
Boston University
crovella@bu.edu

ABSTRACT

Online media is increasingly selected and filtered by recommendation engines. YouTube is one of the most significant sources of socially-generated information, and as such its recommendation policies are important to understand. Because of YouTube's revenue model, the nature of its recommendation policies is fairly opaque. Hence, we present an empirical exploration of the nature of YouTube recommendations, concentrating on socially-impactful dimensions. First, we confirm that YouTube's recommendations generally "lead away" from reliable information sources, with a tendency to direct users over time toward video channels exposing extreme and unscientific viewpoints. Second, we show that there is a fundamental tension between user privacy and extreme recommendations. We show that in general, users who seek privacy by keeping personal information hidden, receive much more extreme and unreliable recommendations from the YouTube engine. This drawback of user privacy in the presence of recommender systems has not been widely appreciated. We quantify this effect along various dimensions, including its dynamics in time, and show that the tradeoff between privacy and unreliability of recommendations is generally pervasive in the YouTube recommendation process.

ACM Reference Format:
Larissa Spinelli and Mark Crovella. 2020. How YouTube Leads Privacy-Seeking Users Away from Reliable Information. In *28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct), July 14–17, 2020, Genoa, Italy*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3386392.3399566

1 INTRODUCTION

Currently, much of the information accessed online is mediated by some kind of recommender system. The increasing and widespread use of recommendation systems has raised concern about how possible biases existing in recommendations can impact worldwide information and public opinion formation.

As a result, research has begun to investigate how personalization can impact the nature of information that is accessed by individuals. One concern is the narrowing of information diversity through the creation of 'filter bubbles' [6, 11, 12, 17]. More recently, the increasing proliferation of unreliable information [5, 20], especially on social media, has been adding a new dimension to recommender system social impact and has been increasing the importance of understanding recommendation policies used on such platforms.

Social platforms such as Facebook and YouTube optimize their recommendations to maximize engagement, while commercial platforms such as Amazon seek to drive purchases. However, although most recommendation algorithms are designed for value-neutral objectives such as engagement and commerce, the resulting recommendations can potentially promote content that is factually unreliable or socially harmful. In this regard, the popular press has recently exposed odd behavior of the Amazon and YouTube recommender systems, including promoting radical, extreme, or unreliable content [7, 9, 13, 14, 19].

YouTube is one of the most significant sources of socially-generated information globally, with over 1.9 billion logged-in visitors each month and more than a billion hours of video watched every day [3]. However, because of YouTube's revenue model, the nature of its recommendation policies is fairly opaque.

In this paper, we seek to move beyond the anecdotal descriptions in the popular press and study the nature of YouTube recommendations quantitatively. We study YouTube recommendations empirically, focusing on socially-impactful dimensions – particularly, recommendations for reliable versus unreliable information sources. To this end, we design and implement a data collection framework to simulate users watching a sequence of recommended videos on YouTube under various experimental conditions. We then classify the channels from the recommended videos in terms of the reliability of their content. Finally, we analyze the empirical results to quantify the extent to which YouTube recommendations shift users away from reliable towards unreliable and even extreme content.

Recommender systems are successful to the extent that they can employ information about users, allowing recommendations to be personalized. At the same time, many users seek to protect the privacy of their personal information while online. Hence, one of the central issues we explore in this paper is the tension between privacy and the nature of recommendations. To that end, our experimental conditions vary in the degree of privacy that our simulated users employ.

Our first contribution is to *quantitatively* demonstrate how YouTube's recommendations generally "lead away" from reliable information sources, including a tendency to direct users over time toward video channels espousing extreme or unscientific viewpoints. By quantifying this effect, we demonstrate that in most cases YouTube leads users away from reliable information very

UMAP '20 Adjunct, July 14–17, 2020, Genoa, Italy

ACM ISBN 978-1-4503-7950-2/20/07…$15.00
https://doi.org/10.1145/3386392.3399566

quickly. That is, most of the change in the reliability of information takes place within the first few recommendations provided by YouTube.

Our second contribution is to measure the effect of user privacy on YouTube recommendations. While many users may consider privacy desirable, we show that protecting privacy has a major drawback: it drastically increases the "leading away" effect of YouTube recommendations. We show that the increase in the proportion of unreliable content increases by a factor of 2× to 3× for users who preserve their privacy while viewing videos. We quantify this effect along various dimensions, including its dynamics in time, and show how pervasive the tradeoff between privacy and unreliability of recommendations is in the YouTube recommendation process.

Finally, we dive into specific questions designed to explore the robustness of these contributions. We examine how the "leading away" effect depends on the specific topic being explored by the user, showing that "leading away" takes place for most of the topics we study, although to varying degrees. We also show that the widely publicized changes made by YouTube to their recommendation policies in January 2019 decreased but did not eliminate the "leading away" effect.

2 METHODS

2.1 Data Collection

As mentioned above, we study YouTube's recommendation strategies by following *chains* of recommendations made by YouTube. Starting from a specific *search query*, we simulate a user who watches the resulting video and then selects one of YouTube's recommendations to watch next. Each chain is collected under a specific *privacy scenario*, and the next video to watch in each case is selected from the list of recommendations according to a *video selection* strategy. We explain each of these experimental aspects in the following subsections.

2.1.1 Privacy scenarios. In order to explore how YouTube recommendations change as a function of what user features are visible to YouTube, we consider four privacy scenarios:

Logged. The user identity is exposed by being logged into a Google account. We used a single university-provided account.

Normal. The user has not logged into a Google account, but uses normal browsing mode which could be potentially tracked by cookies.

Private. The user has not logged into a Google account and uses a private browser session that disables cookie placement.

Tor. The user has not logged into a Google account and uses a private session in a Tor-enabled browser that obfuscates the user's IP address by passing through the Tor network.

2.1.2 Search queries. Each collected video chain starts with a search query. For these queries, we use the top 10 News Google Searches of 2017 in the United States [2]. We choose these because they represent a set of queries that would be likely as starting points for watching videos on YouTube. The search queries used were: *Hurricane Irma*; *Las Vegas shooting*; *Solar Eclipse*; *Hurricane Harvey*; *Bitcoin Price*; *North Korea*; *Hurricane Jose*; *Hurricane Maria*; *April the Giraffe*; and *DACA*.

(a) Trustable **(b) Neutral** **(c) Extreme**

Figure 1: Channel Name Word Clouds by Classification.

2.1.3 Video Selection. YouTube's recommendations are provided in a list on the right side of the screen while a video is playing, which we call the *recommedation list*. An important aspect of our experiment is the choice of how the next video to be watched is selected from this list (*video selection*.)

"Auto-play" mode in YouTube simply plays the top item in the recommendation list. This mode is the default behavior of YouTube and is followed when no other user action is made. Hence, we treat the top item as the one most strongly recommended. This video selection strategy is *top item*.

In contrast, to understand the impact of the recommendations rank, we consider a strategy in which the video with the lowest ranking is the one chosen for viewing next. We refer to this video selection strategy as *bottom item*.

We refer to a sequence of videos watched in this way after a single query as a *chain*. In our video selection executions we avoid video repetition inside a chain. Then, if the selected video to play next had already played in the current chain we choose instead the highest unplayed video (or lowest unplayed video) when the video selection is *top item* (*bottom item*).

2.1.4 Features collected. Each YouTube video is published by a YouTube channel. Channels are either based on Google personal accounts or Google brand accounts. Channels with more than 100,000 subscribers that belong to an established creator or are the official channel of a brand, business, or organization can receive a YouTube verification badge checkmark upon request [1].

For each video viewed during our experiments, we collect a set of features. Features are either derived from the video or its channel. The video features we collect are: the current date, the video publication date, the video number of views, the video number of likes, the video number of dislikes, the video number of comments, the video title, the video duration, the video description, the video content category (selected on video publication), the channel identifier, the channel title, the channel number of subscribers and, the channel verification badge status.

2.1.5 Data Collection Process. Putting all the above parts together, the overall structure of our data collection process is given in Algorithm 1. The framework is implemented in python and uses Selenium to simulate user behavior.

2.2 Classification

We classify each recommended video according to its channel. We place each channel into one of three categories: *trustable*, *neutral* or *extreme*. Each channel encountered in our data was classified manually. Manual classification was done via inspection of the most popular movies of the channel as well as the channel description. The criteria we used for channel classification are:

Algorithm 1 YouTube Data Collection

Require: Privacy scenario, Search term, Selection Strategy
1: Perform a search query and get its recommendation list.
2: **repeat**
3: Select video from list according to selection strategy.
4: **if** advertisement appears **then**
5: Wait for the end of the ad or skip it.
6: **end if**
7: Watch selected video for up to 5 minutes of elapsed time
8: collect data about the video and its channel
9: get new recommendation list
10: **until** video chain reach 20 videos

Trustable. Channels identified as trustable are channels from established news sources. Most trustable channels are run by news sources from television, or are credible scientific channels that provide content with externally checkable references.

Extreme. Channels that are identified as extreme are those that have content that deny established scientific knowledge, incite hate or promote fake news.

Neutral. Neutral channels are all other channels – those that are neither trustable neither extreme.

In Figure 1 we illustrate of the type of channel in each classification using various word clouds based on channel names. Figure 1a shows a word cloud of *trustable* channel names, displaying traditional news sources such as *ABC*, *CBS*, *Fox*, and *CNN*. Figure 1b shows a word cloud of *neutral* channel names, and is dominated by entertainment – music and gaming channels – such as *YoungBoy Never Broke* and *TmarTn2*. The word cloud of *extreme* channel names, in Figure 1c, shows that extreme video channels use attention-getting names such as "True", "Mysteries" and "Top".[1]

2.3 YouTube Recommendations

During the period of our study, YouTube generated recommendations using a deep neural network that implements a two-stage approach of candidate generation followed by ranking [10]. This approach was designed to deliver high performance on key metrics: precision and increased watch time, while handling the challenges of scale, freshness and noise.

YouTube has stated that is continuously working on improving its search results and recommendations using user feedback, external evaluators trained using Google public guidelines (that evaluate the content quality and publisher reputation) and other signals. Relevant to our study, in January of 2019 YouTube announced on its blog that it would reduce recommendation of borderline content and content that could misinform users in harmful ways [4]. This fell in the middle of our study period, an event that we analyze in detail in Section 4.5.

3 DATASET

Our dataset consists of a set YouTube chains collected (as specified in Section 2.1) between October 2018 and April 2019. Each experimental setting was replicated 256 times, with one chain collected each time. As a result, the dataset consists of 4 (privacy scenarios) × 2 (selection strategies) × 256 (chains) × 20 (videos per chain) = 40,960 videos. There were 25,091 unique videos in this set. The 10 search queries were evenly distributed across replications.

[1]To allow readers to examine our channel classifications in detail, as well as to reproduce our results, all data and code in the form of Python notebooks will be released upon paper publication.

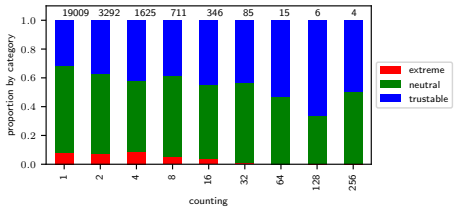


Figure 2: Binned counts of the number of appearances of each video, and classification of videos in each bin.

Although our experiments use a small set of search queries, the videos and channels that are collected are broadly distributed. We show this in Figure 2, which bins videos according to how many times each appeared in the dataset. Each bar shows the distribution of classification for the videos in that group, and numbers at the tops of bars show how many videos fall in each group.

We note that most of the videos appear only a few times. For instance, 46.4% of the videos were recommended just once, and only 1.8% of the videos were recommended 16 or more times. The figure also shows that videos belonging to *extreme* channels received relatively few recommendations overall, while videos pertaining to *trustable* channels receive more recommendations – which can be observed by the increasing proportion of this classification for higher-count bins.

4 RESULTS

In this section we present the results of our analysis of the effect of YouTube recommendations on the reliability of content seen by users. We present reliability classification in *ternary plots*. In a ternary plot, each side of the triangle represents one of the three classification types and each point within the triangle represents a particular proportion across the classifications. We denote this proportion across the classification as the proportion mix. The axis values grow counter-clockwise. For any point, the corresponding fraction values for each classification type can be obtained by the projection of this point into the corresponding axis. Point projection follows a parallel line to the triangle side where this axis has value 0 (clockwise side), for instance, the *neutral* projection line is parallel to the *extreme* triangle side.

4.1 YouTube recommendations lead away from trustable sources

Our first result shows that YouTube's recommendations guide users toward less-reliable sources over time, in two ways: (a) they guide users away from trustable sources, and (b) they increase the fraction of extreme content recommended to the user. To demonstrate this, in Figure 3 we show the characteristics of the sequence of recommended videos, aggregating all privacy scenarios and using the *top item* video selection.

The shift away from reliable videos can be seen in Figure 3. In this initial ternary plot, we use a heatmap overlaid on the plot to signify overall degree of extreme content. The figure shows that at

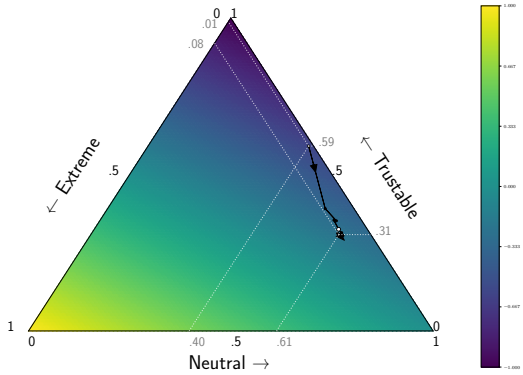
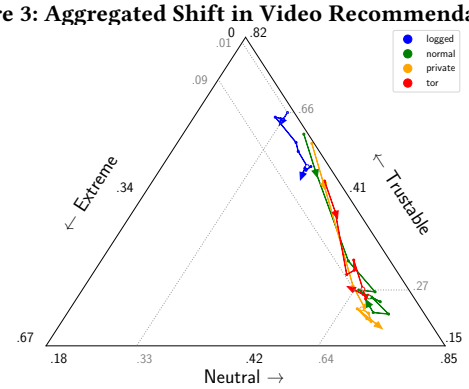
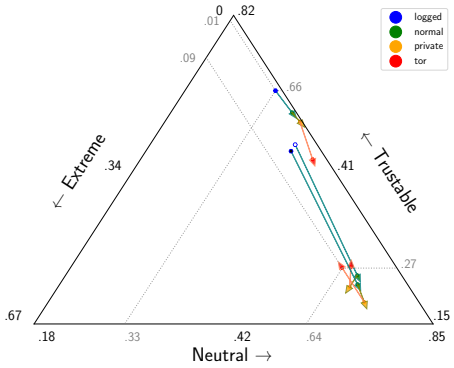


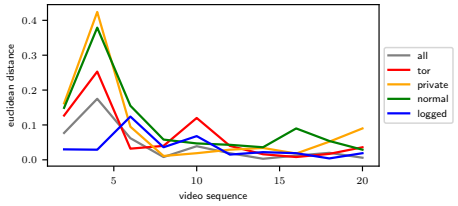
Figure 3: Aggregated Shift in Video Recommendations.



(a) classification proportions shifts



(b) progression path of increases in privacy (*logged* → *normal* → *private* → *tor*) for three time samples (initial, middle and end of chain)



(c) Recommendation Shift per Privacy Set.

Figure 4: Impact of Privacy on Shift in Recommendations.

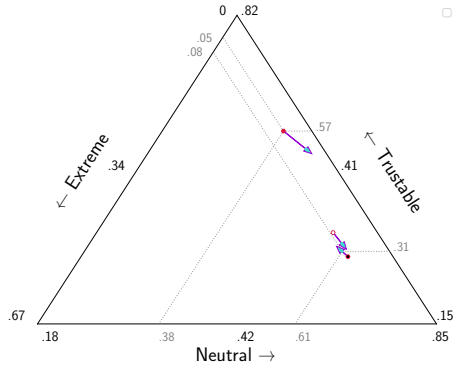


Figure 5: Impact of Video Selection Strategy (*bottom item* → *top item*) in time (initial, middle and end of chain)

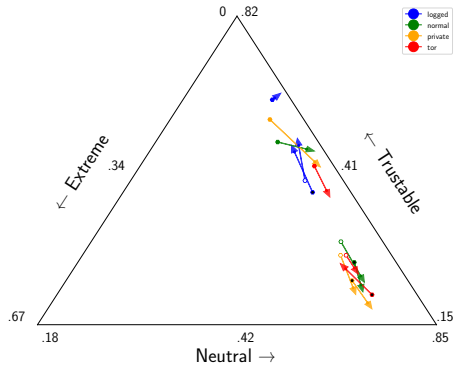


Figure 6: Impact of Video Selection Strategy (*bottom item* → *top item*) per Privacy Scenario.

the beginning of the sequence, 59% of videos come from trustable channels, while at the end of the sequence, only 31% of videos come from trustable channels. Furthermore, the fraction of extreme videos increases over by more than a factor of six, from 1.2% to 8%.

Examining the trajectory in Figure 3 shows an important observation: the movement away from reliable videos is initially very fast, after which the change comes more slowly. In the figure, dots and arrows indicate the sequence progression inside a video sequence: the first and last sequence points are represented by arrows, and the middle of the sequence is represented by an open circle. In this way, the length of the lines between points or arrows represents how much the fractions have changed within two observations. Using these, we observe that most of the changes in the proportion mix occur in the first half of the observations and changes in the second half tend to be generally smaller. This suggests that much of the significant changes in proportion mix occurs as a result of the initial recommendations, and implies that trace measurements longer than 20 steps would not likely to show vastly different results in terms of the final proportion mix.

4.2 Private users get less reliable recommendations

The results in the previous section are aggregated over all privacy scenarios, and hide important differences. In fact, the tendency

for YouTube recommendations to lead away from reliable sources depends enormously on the user's privacy settings. We find that *privacy-seeking users are much more likely to be directed away from reliable sources and toward extreme videos.* We show this effect in Figure 4a where the proportions mix for each privacy scenario (*logged*, *normal*, *private* and *tor*) is shown in the ternary plot in blue, green, orange and red, respectively. We observe that all privacy scenarios show a decline in the fraction of trustable channels over the recommendation sequence. However the four privacy settings have considerable and important differences.

We note first of all that the overall effect of YouTube recommendations is much weaker when users are logged in – the chain path for *logged* users is much shorter than the others. Furthermore, logged-in users are those with the largest proportion of trustable channels initially recommended, and the smallest decline in the fraction of trustable channels over time. For *logged-in* users, there is a difference of 13.1% in the trustable proportion from the initial to the end observation, while for the other privacy scenarios this difference is much larger (30.1% for *tor*, 37.3% for *normal*, and 39.3% for *private*).

Second, we see that *tor*, *normal* and *private* settings tend to arrive at nearby endpoints, with relatively low fractions of trustable channels and high fractions of extreme channels. However, the initial recommendations provided by YouTube are quite different for *tor*, compared to *normal* and *private*. The privacy scenario *tor* starts with a low fraction of trustable channels (50.8%) while all the other settings have more than 59% of trustable channels in their initial recommendations. The lack of significant difference between *normal* and *private* suggests that if a user is not logged in, then private browsing versus normal browsing has little effect on the YouTube recommendations. This may reflect aspects of how YouTube identifies users during a browsing session.

It's important to note that the phenomenon seen in the combined data, in which the proportion mix changes fast during the first few recommendations but slower later on, is present in each individual privacy scenario as well. Figure 4c measures this effect by presenting the euclidian distance between ternary observation points. In this plot, we can observe that the most significant differences are between the first observations – with distances approximately ten times larger than the end observations. This trend confirms that our conclusions regarding privacy scenarios would not likely change by observing more extended sequences of videos.

As noted, when privacy increases, there is a decrease in recommendations from reliable sources, and an increase the fraction of extreme channels. To measure this effect, in Figure 4b we show paths that progress along increases in privacy: from *logged*, to *normal*, to *private* to *tor*. Each path corresponds to the same time in a chain: either the initial recommendation, or the sequence midpoint (10th recommendation), or the sequence endpoint (20th recommendation). The initial points are filled, the sequence midpoints are open circles, and the sequence final points are black circle.

Figure 4b shows that early in the recommendation sequence, an increase in privacy increases the amount of extreme channels. However, for the middle and end values, there are similar amounts of extreme videos among the three privacy scenarios that doesn't disclose the user identity. Initially, the most significant increase is between *private browsing* and *tor browsing*, however by the sequence midpoints, the largest difference is between *logged* and *normal*. We also note that by the end of the recommendation sequence, the main shift in going from *logged*/*normal* to*normal*/*private* is an increase in the fraction of extreme channels.

Furthermore, the difference among *tor* and *private* reveals the role played by the IP address obfuscation. Our results show that IP address is an important factor in the initial recommendations – marked by the long arrows – but it loses its impact over time – marked by the short arrows in the middle and end observation. Also, the difference between *private* and *normal* suggests that when cookies are disabled, there is a slight increase in user exposure to more extreme videos. Finally, the difference between *normal* and *logged* reveals the possible impact of knowing the user identity on YouTube's recommendations. In our results, this knowledge has a notable impact on YouTube's recommendations and significantly minimizes the exposure of the user to more extreme content.

4.3 YouTube more strongly recommends less reliable sources

Although the sequence of recommended channels tends away from trusted sources and tends toward extreme sources over time, we would like to assess how important the particular recommendation methods used by YouTube are to this effect. For example, it is possible that simply recommending a random set of related videos would move the user away from trusted sources.

To gauge this effect, we look at the differences between video selection of *top item* and *bottom item*. If the YouTube algorithm is actively favoring unreliable channels then we will see a greater tendency away from reliable channels when following the *top item* as compared to the *bottom item* in each recommendation list. In fact, we show that the YouTube recommender system is influencing the video outcome, that this influence is stronger in the inital part of the video sequence, and that indeed, *the recommender system is leading users to more extreme channel sources.*

In Figure 5 the lines and the arrows represent the difference between the initial, middle and end observation time among following the *bottom item* and the *top item*. From the observation of *bottom item* choice, the initial, middle and end point is marked respectively by a full colored circle, a white circle with color border, a black circle with color border. First, note that the fact there is a difference between following the top or following the bottom recommendation indicates that the recommender system is actively working and not just suggesting random videos. Second, the length of the lines connecting sequences following the top and the bottom is decreasing over time. For instance, the initial arrow length – computed by the Euclidean distance between the projected points – is .049 while the end arrow length is .013. This length reduction indicates that the impact of the recommender system is considerably stronger in the initial recommendations.

Finally, the arrow direction from bottom recommendation to top recommendation indicates the prioritization of the recommender system. These directions show that the recommender system actively shifts the proportions of videos away from trustable channels in the initial and middle observation. This shift represents a reduction of 3.9% in the fraction of trustable channels for the initial

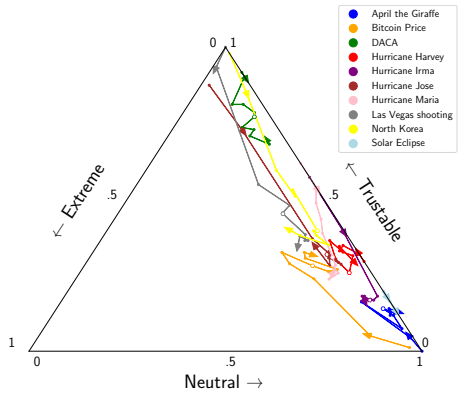


Figure 7: Recommendation Shift by Search Query.

(a) April the Giraffe **(b) DACA** **(c) Solar Eclipse**

Figure 8: Word Cloud of Channels' Names by Search Query.

observation, 2.4% for the middle observation and an increase of 0.4% of extreme channels for the final observation.

Importantly, however, this effect is not the same for all privacy settings. While the YouTube recommender system guides privacy-seeking users towards less reliable sources, when the user identity is revealed (*logged* users) the system can in fact favor more reliable sources. This effect is shown in Figure 6. In that plot arrows indicate the shift from following the bottom toward following the top recommendation; colors show the privacy setting adopted; and the initial, the middle and the final observations are marked as before. First, note that the initial and middle points arrows of *logged* (in blue) – with respectively length of .035 and .059 – are shorter than the other settings ones – which arrows lengths range from .002 up to .1. This shows that the recommender system has less effect on users whose identity was revealed. Second, observe that only for this setting, *logged*, the system effect manifest a shift towards *more trustable* and less extreme videos. For instance, in the middle observation point, there is for *logged* an increase of trustable of 7.6% and a decrease of extreme of 2.7% while the other settings for the same observation point show a a decrease of trustable of 7.4% for *normal*, 7.1% for *private* and 2.8% for *tor*. Thus, we find that the YouTube recommender system, while clearly leading privacy-seeking users away from reliable and towards extreme videos, does not have the same effect for users whose identity is known to the system.

4.4 YouTube's recommendation effect varies depending on topic

Next, we show that YouTube's recommendation system does not affect all query topics equally. To illustrate this phenomenon, we show in Figure 7 the ternary plot over time for each query. The figure shows that the effect of YouTube recommendations varies considerably for different topics.

First of all, for most queries, YouTube leads users away from reliable information toward unreliable and extreme content. This

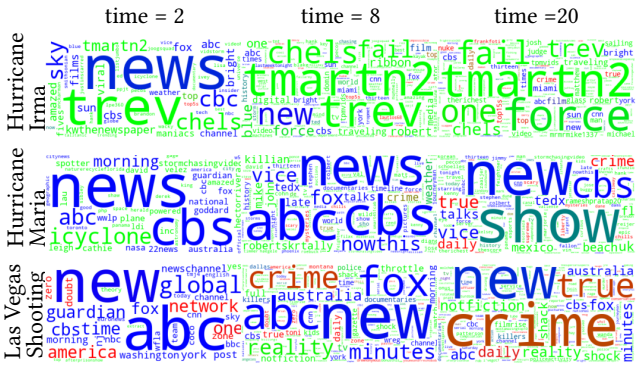
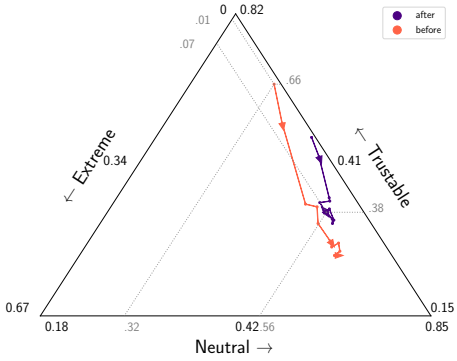
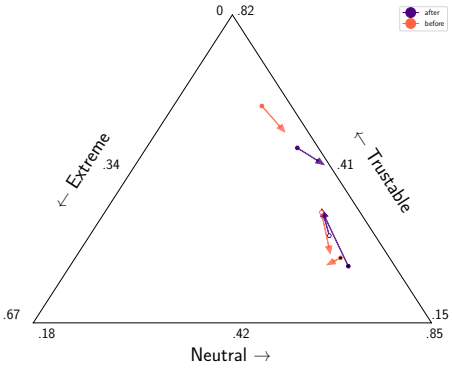


Figure 9: Word Cloud of Channels' Names by Search in Time.



(a) classification proportions shifts

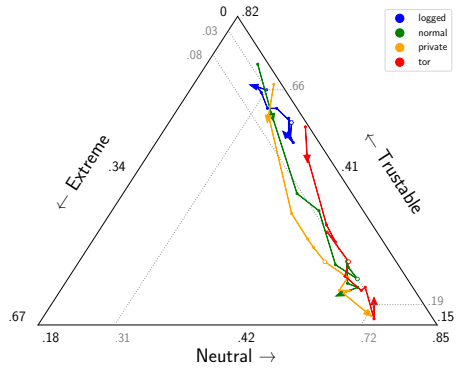


(b) video selection strategy difference (*bottom item* → *top item*) in time (initial, middle and end observation)

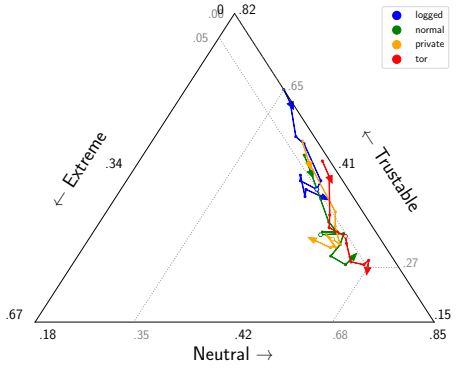
Figure 10: Recommendation Shift Before and After YouTube Policy Change.

is consistent with our results above. However, for some queries (*April the Giraffe* and *Bitcoin Price*) the overall movement is toward a mix of more reliable, but also more extreme content. Second, some queries are very strongly affected by YouTube recommendations (*Las Vegas Shooting* and *North Korea*) while other queries are not strongly affected (*Solar Eclipse*). In fact, the path length for the former is almost 20 times longer than that of the latter.

Inspecting the word clouds for queries we can shed light on the reasons for the differences we observe. Figure 8 shows the overall word clouds color coded, with word color proportional to the classification of the channel that the word comes from.



(a) before



(b) after

Figure 11: Recommendation Shift Before and After YouTube Policy Change, by Privacy Scenario.

Figure 8 shows why some queries are relatively unaffected by the YouTube recommendation system. The small changes for queries such as *DACA* or *April the Giraffe* can be understood because the former is hard news and covered mostly by traditional news channels such as *CNN* or *ABC* while the latter is soft news covered mostly by animal-related channels (classified as neutral). Additionally, the small effect for the query *Solar Eclipse* is explained by an ambiguous results split between hard news (news about the 2017 Solar Eclipse) and soft news (the title of a song, "Solar Eclipse" by the popular singer "YoungBoy Never Broke Again").

On the other hand, some topics are much more strongly affected by the YouTube recommendation system. To unravel the reasons for different length paths, Figure 9 presents word clouds of specific points in the sequence for some of the queries, where each word is presented in a color related to its classification category fractions – red for *extreme*, green for *neutral* and blue for *trustable*. First, *Las Vegas Shooting* is the longest path of Figure 7 and we can observe in the time clouds of Figure 9 its initial dominance by traditional news channels that are replaced over time with more neutral channels and conspiracy channels (here classified as extreme). Comparing the paths of the queries *Hurricane Irma* and *Hurricane Maria* in Figure 7, we note that the latter is much more strongly affected than the former. Referring to Figure 9 , we see that *Hurricane Irma* recommendations are dominated by entertainment channels late in its sequence, shifted by a popularity bias caused by a hurricane video posted by popular YouTuber gamer "Tmartn2". However, *Hurricane Maria* presents a more balanced path between entertainment and traditional news.

4.5 YouTube's policy change did not fully remove the shift toward unreliable sources

As we noted above, during our data collection period YouTube implemented a change to its recommendation policy. Our analysis shows that after that change, in late January of 2019, YouTube was still leading users away from reliable sources over time. However, there was a reduction in the tendency to extreme recommendations. To illustrate that change we separate our data into two parts: data collected before February and data collected from February onward – sampling the data for the top item selection strategy, in a way that both portions of the dataset have the same number of experiments per privacy scenario and query.

The change in how YouTube leads users away from realiable sources before and after the policy change is expressed in Figure 10a. Comparing trajectory paths we can see that in both cases, users are led away from reliable sources and toward neutral and extreme content. However, after the policy change the effect of the recommendation system is decreased overall, and fewer extreme channels are recommended. For instance the before path has an increase of 8.5% in the fraction of extreme recommendations (going from 1.2% to 9.7%) while the path after has an increase of 5.9% (going from 0.6% to 6.5%).

Comparing the effect of the recommender system before and after the policy change we find that for initial recommendations the effect was similar, favoring untrustable sources, however for later recommendations the revised recommender system shifts toward promoting more trustable channels. We see this in Figure 10b, which has an arrow connecting the values from the bottom item to the top item selection strategy for the initial, mid-point and end-point (marked respectively with full colored circle, a white circle with color border and a black circle with color border), comparing before and after the policy change. In this figure, while all the observed points before the change show a shift away from reliable sources, the mid and end point after the change show a tendency toward favoring more trustable sources.

Finally, we can observe in Figure 11 that the overall trend regarding the tradeoff of privacy and extreme content exposure is likewise present both before and after the YouTube policy change.

5 RELATED WORK

Although previous work has addressed how recommender systems can impact information access by creating "filter bubbles" or "echo chambers," relatively few studies have explored the impact of recommender systems related to the reliability of content. In this section, we contrast our work with studies that have focused on YouTube and its recommendations.

Our study takes inspiration from recognition in the popular press that YouTube's recommendations can lead to extreme content. Articles such as [19] and [14] describe the radicalization of YouTube recommendations and discuss the social implications of this effect. The article [7] continues the discussion and presents some of the actions taken by YouTube in response to its critics. These articles provide essential context for our study by highlighting issues, but none performs a quantitative analysis of YouTube's recommendation system. In contrast, we quantify the strength and dynamics

of YouTube's "leading away" effect, showing that most of its effect takes place within a sequence of just a few recommendations.

Our work shares some similarities with [15], which also performs an empirical exploration of YouTube recommendations. However, that work looks at a much smaller dataset with a simpler overall experimental design; by studying a larger dataset we provide greater robustness of results. Most importantly, it does not explore the trade-off between recommendation properties and privacy, nor does it analyze time dynamics, the impact of YouTube policy changes, nor the relationship to query topic.

The authors in [16] investigate the recommendation of extreme-right videos on YouTube by using a content categorization schema. Like our study, that work notes how quickly the YouTube recommender system can deviate from reliable content. However, that work focuses on one specific niche of the content spectrum (extreme-right) and the discovery of its ideological bubbles. Our work adopts a more extensive notion of extreme content and consequently provides a broader understanding of the "lead away" effect of YouTube recommendations. Further, other work [8, 18] also covers the discovery of ideological bubbles with harmful social impact on YouTube, but without examining YouTube's recommendations.

6 DISCUSSION AND LIMITATIONS

An important issue in our study concerns the way that we have defined the three categories of channels, and the related policy implications. A main observation in our study is that YouTube shifts its recommendations over time from 'reliable' to 'neutral' channels. We note however that this sort of shift is not necessarily undesirable in general. Indeed, it can be argued that an important goal of recommendation systems in a system like YouTube is to expose users to the "long tail" of content that lies outside of the realm of mass media. This benefits users by diversifying their influences, and it provides opportunities for lesser-known producers to build an audience. Hence without further study we cannot conclude that the shift from 'reliable' to 'neutral' channels is socially undesirable.

However, our definitions for 'extreme' content are intended to specifically capture socially undesirable content. And for that category, we find that the trends we identify – with respect to time, privacy, strength of recommendation, and topic – are all consistent with our high level conclusions; and that in many cases the fraction of 'extreme' content increases by a factor of 6× to 8× from the beginning to the end of a recommendation sequence.

7 CONCLUSION

In this paper we have presented an empirical exploration of the nature of YouTube recommendations. We developed a data collection framework that impersonates users watching videos on YouTube, for different privacy scenarios and video selection policies. We classified the pool of recommended channels and quantified changes to nature of the recommended content over time.

Our results show that YouTube's recommendations typically lead users away from reliable sources over time. Importantly, we pointed out where in time this shift happens, demonstrating how quickly users can be exposed to extreme information. A particular focus of our study is the tension between user privacy and extreme recommendations, and we expose the fact that privacy-seeking users are much more likely to be led away from reliable sources and towards extreme videos. Then, we show how YouTube's "lead away" effect varies according to the query topic, but that most topics we studied exhibit the effect. Finally, we find that the last changes in the YouTube recommendation policy have reduced but not yet solved the "lead away" effect.

Nonetheless, our results suggest that engagement-driven recommendations, such as used by YouTube, can have undesirable interaction with privacy-seeking users, resulting in a tendency to strongly direct such users toward unreliable information. Taken in the context of the currently-dominant business model of advertising-supported content publication, the ongoing evaluation of these effects is vital for understanding their impact on society.

Acknowledgments. This material is based upon work supported by the National Science Foundation under grant numbers IIS-1421759 and CNS-1618207. We thank the anonymous reviewers whose comments improved this paper.

REFERENCES

[1] 2019. Verification badges on channels. (2019). https://support.google.com/youtube/answer/3046484?hl=en Accessed: 2019-04-20.
[2] 2019. Year in Search 2017: See what was trending in 2017 - United States. (2019). https://trends.google.com/trends/yis/2017/US/ Accessed: 2018-04-30.
[3] 2019. YouTube for Press: YouTube in numbers. (2019). https://www.youtube.com/intl/en-GB/yt/about/press/ Accessed: 2019-04-30.
[4] 2019. YouTube Offical Blog: Continuing our work to improve recommendations on YouTube. (2019). https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html Accessed: 2019-04-30.
[5] Jason Reifler Andrew Guess, Brendan Nyhan and. 2018. Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign. (2018).
[6] Eytan Bakshy, Solomon Messing, and Lada Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* (07 May 2015).
[7] Mark Bergen. 2019. YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant. (2019). https://www.bloomberg.com
[8] A. Bermingham, M. Conway, L. McInerney, N. O'Hare, and A. F. Smeaton. 2009. Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation. In *2009 International Conference on Advances in Social Network Analysis and Mining*. 231–236.
[9] Guillaume Chaslot. 2019. The Toxic Potential of YouTube's Feedback Loop. *Wired* (2019). https://www.wired.com
[10] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 191–198.
[11] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring Personalization of Web Search. In *Proceedings of WWW*. ACM.
[12] Huyen Le, Raven Maragh, Brian Ekdale, Andrew High, Timothy Havens, and Zubair Shafiq. 2019. Measuring Political Personalization of Google News Search. In *Proceedings of World Wide Web '19*. ACM, New York, NY, USA.
[13] Kalev Leetaru. 2019. Do We Need New Laws Forcing The External Auditing Of Social Media Algorithms? *Forbes* (2019). https://www.forbes.com/
[14] Paul Lewis. 2018. 'Fiction is outperforming reality': how YouTube's algorithm distorts truth. *The Guardian* (2018). https://www.theguardian.com
[15] Jack Nicas. 2018. How YouTube Drives People to the Internet's Darkest Corner. *The Wall Street Journal* (2018). https://www.wsj.com/
[16] Derek O'Callaghan, Derek Greene, Maura Conway, Joe Carthy, and P$#225;draig Cunningham. 2015. Down the White Rabbit Hole. *Soc. Sci. Comput. Rev.* 33, 4 (Aug. 2015), 459–478. DOI: https://doi.org/10.1177/0894439314555329
[17] E. Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin Group USA.
[18] Ashish Sureka, Ponnurangam Kumaraguru, Atul Goyal, and Sidharth Chhabra. 2010. Mining YouTube to Discover Extremist Videos, Users and Hidden Communities. In *Information Retrieval Technology*. Springer Berlin Heidelberg.
[19] Zeynep Tufekci. 2018. YouTube, the Great Radicalizer. *New York Times* (2018), SR6. https://nyti.ms/2GeUCDY
[20] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.