AMENDED Exhibit 507

PLAINTIFFS' OMNIBUS OPPOSITION TO DEFENDANTS' MOTIONS FOR SUMMARY JUDGMENT

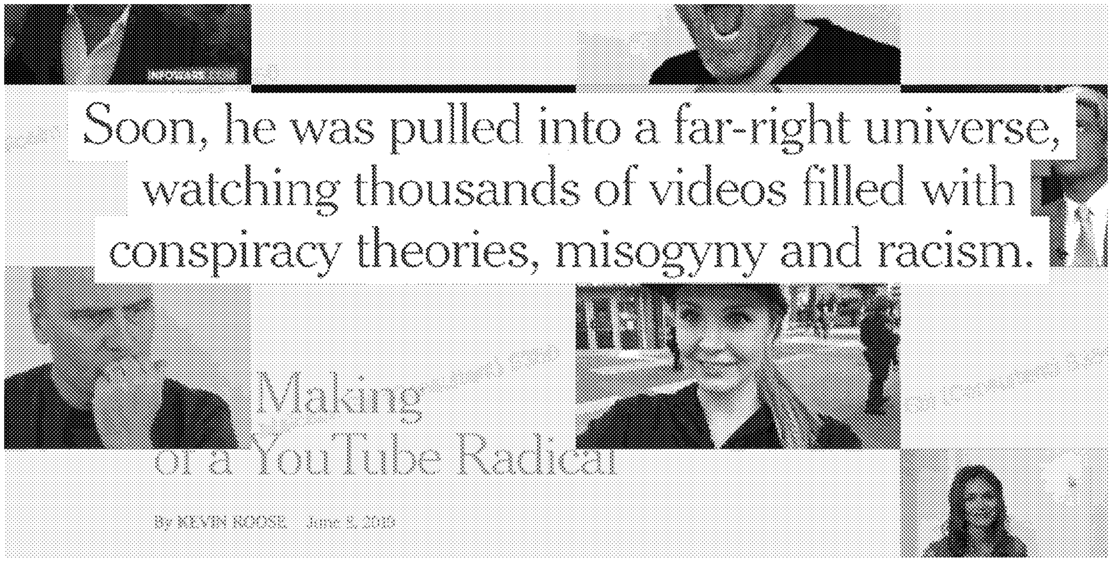
Rabbit Holes ("Project Bubble") Presentation (P&C)

7/28/2020

Rabbit Holes

General

Social media rabbit holes gained prominence with 2018 New York Times coverage of YouTube's content recommendation system:



A rabbit hole occurs when a content recommendation system leads a user to new types of content based engagement behavior. A user engages with the new content genre more and more until it becomes the only, or majority of content consumed.

Rabbit Holes on TikTok

User Feed Examples

SafetyVideo_ForBen_V2.mp4

- 20 sequential videos from a real user's FYP

- 12 sequential videos from a real user's FYP

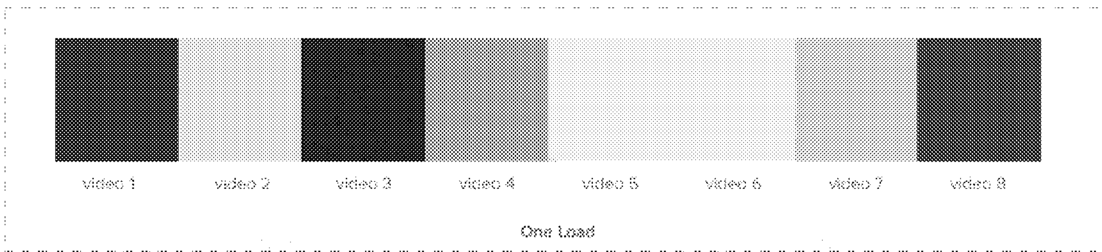
Trust & Safety: Why We Care

Rabbit holes raise trust and safety concerns for the following reasons:

1. Leads users to harmful content they would not have encountered otherwise.
2. There will always be grey-area policy cases. On a one-off basis they pose little risk of harm, but that's not the case if all grey-area cases of a particular type of content are concentrated on a user's feed.
3. Users could encounter violative mismoderated content in a high concentration
4. R2 policies (conspiracy theory, stereotypes, some sexualized content policies) are enforced at 12,000vvs. R2 violative low vv content could be concentrated on a user's feed before it hits moderation.

A Brief Introduction to Recommendation

When a new user downloads and installs TikTok, when they first open the app, a request is sent to the recommendation system for a "first load of videos". This load contains 8 videos and these 8 videos are chosen from a small pool of videos that are known to be very popular on TikTok and that have also passed safety moderation queues so we have verified they are appropriate for a general audience:

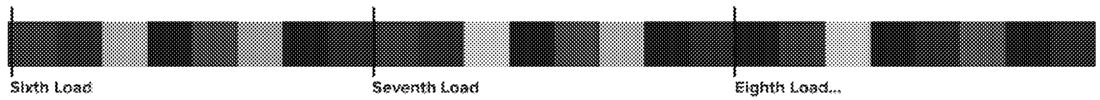


As the user watches these videos, the algorithm records all of the user's interactions with the videos, for example, did the user do any of the following:

- Like, share, comment, follow, finish, skip, etc...

Once the user finishes this first load of 8 videos, a new load of 8 videos is requested from the recommendation system. The same process ensues throughout the lifetime of the user and their "feed" can be said to be made up of many, many loads. For example, a user who watched 1,000 videos, has requested 125 loads from the recommendation system:

As a user finds an interest(s), the recommendation system begins to recommend more and more "similar" videos, and their feed can be said to have a specific "feel" to it:



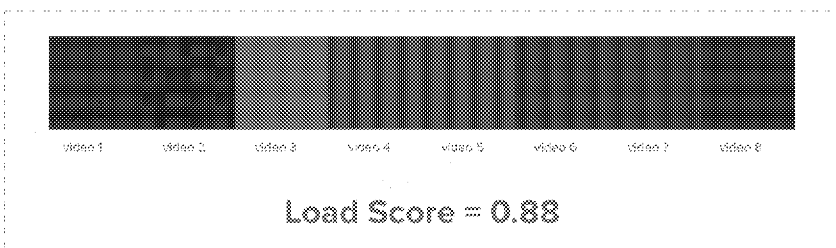
To make sure that a user's feed does not become too "similar" and that there remains an element of "diversity" within the user's feed, we calculate a "load score", translating to the overall similarity of the 8 videos in the load. The similarity is calculated by comparing the recommendation score for each video, and how close these scores are to each other, returning a number between 0-1 (0 representing zero similarity and 1 representing perfect similarity, or 100% similarity meaning it's the same video). We will call a load highly similar if its score is >= 60%. By calculating "load scores", we can measure the following:

- # users who have experienced at least one highly similar load
- % of DAU to have at least one highly similar load per day
- # loads that are highly similar per day
- % of loads that are highly similar per day
- Average time (days or # of videos) for a user to reach their first highly similar load
- etc...

Utilizing the load score can help us automatically identify potential rabbit holes because we can return all of the users on a given day who had at least one highly similar score and review the recent watch list of these users to see if they have a feed of highly similar videos.

Identifying Bubbles

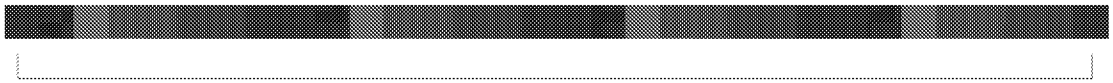
Everyday, we can generate a report of users who encountered a highly similar load (>=0.60):



Example Load 1: sexualized link

Example Load 2: Qanon related

For users identified with a high load score, we can then pull their recent watch history of 100 most recent videos to see if these recent videos are also highly similar, or of the same topic:



100 Most Recent Videos
Similarity Score = 0.66

User 1 Recent 100: sexualized link

User 2 Recent 100: political related

More examples: Bubble Score Load

Classifying Bubbles

From TnS perspective, we care most about rabbit holes where users are consuming videos with potentially harmful topics. For each highly similar load, we will predict the topic of the load by analyzing the text of the videos (video caption, hashtags, sticker text) to see if the video text in these loads triggers any known watch words ("qanon", "pizzagate", "ww1wga", etc...).

For users whose load hits a trigger word, we will then pull their recent watch 100 to see if their feed also has a high degree of similarity. We can then determine how many users on a daily basis are experiencing potential rabbit holes with risky topics and identify the topics most related to rabbit holes. We can leverage this data to inform moderation policy, algo policy, and product features to decrease the negative impact of repetitive harmful content.

Defining Bubbles

Utilizing the methodology of calculating the similarity of a load, we will define a rabbit hole on TikTok as any user whose recent watch 100 has a similarity score of >= 0.60.

How Does a User get into a Bubble?

If a user continuously shares, likes, comments, and finishes, specific types of videos, these signals will inform the recommendation system that the user likes this topic of video and the recommendation system will recommend more similar videos, increasing the chance that a user will encounter a bubble of similar videos.

The fastest way a user gets into a specific bubble immediately as a new user, is directly searching a specific topic of video, and then clicking and watching these videos from the search page. This type of video will then appear more frequently in the loads of the user's FYP. If the user continues to finish/like/share/comment on these videos, the recommendation system will continue to recommend them. If the similarity score of the load of videos remains unchecked, the user, theoretically, could remain in the bubble forever, if the user does not change their actions against the videos.

This is why we are implementing the load score, so that we can ensure that, for each load, there is a certain threshold of diversity and the videos don't remain too similar. Furthermore, we aim to create actionable metrics around feed and user diversity to improve the richness of the underlying viewing experience.

Example

Bubble 2550

We looked at 127 videos from this user's most recent FYP watch history.

- 73/126 TikToks were conspiracy theory content
- 50/73 TikToks never reached R2 moderation
- 44/73 TikToks were Not Recommended but reached user's feed prior to moderation action
- 7/73 TikToks reached R2 moderation and were mismoderated

Simple examples of how we can steer users out of rabbit holes

There are a variety of ways we can improve the bubble experience on TikTok, including video tagging, algo policy rules, and new product features:

Type	Description	Pros	Cons
Video Tagging	Tag videos with associated risk tags (algorithmically, human, or both).	Can instantly size risky video pool and (i) who is watching them and to what extent and (ii) who is creating them.	Algorithmic labeling is still very difficult and we rely on 10K manual labeling which does not cover these risk tags. Manual labeling requires large human efforts.
Algo Policy Rules	Create rules for each load that specify the max threshold of similarity between the 8 videos.	Can better regulate the probability that a user will enter a sustained bubble.	Could impact the user experience such that feeds do not become as personalized and distinct.
Product Levers	Create product features that aim to guide users out of potential bubbles. Could include a "shuffle" feature or pop-ups ("Is your feed too repetitive? Try switching it up!") or restart buttons ("restart your feed now").	User controls their experience, deciding when they want to exit their bubble.	Users could still choose to remain in harmful bubbles.

Trust & Safety

Bubbles leading users to harmful misinformation are of highest concern ahead of the 2020 US election. Potential solutions from this include:

Type	Description	Pros	Cons
Policy on Feed Diversification	Create a policy with a criteria for 1) concentration, and 2) harm, for when a bubble experience merits intervention.	• New and useful light touch moderation tool, "do not show more than X videos on Y topic to any one user" • Action on content at scale, less human input required	• Problematic content still in the ecosystem • For misinfo, damage could already be done by the time we detect and intervene • What would problematic content be replaced with? How long would it take to build technical capabilities to do this?
Early detection of violative videos	Send any newly created content predicted to be linked to an already identified harmful bubble to moderation before FYP distribution	• Intervention at low vv • Removes violative content from ecosystem	• Bubble has to form before we can use this strategy • Doesn't solve for grey area concentration issue • Human resource intensive, need to create more moderation queues
Changing enqueue logic	Make it so certain R2 policies, like conspiracy theory, are moderated at a lower vv.	• Few resources needed on the technical and human side	• Training R1 moderators on nuanced policy issues

Action Items and Futures

Product and Algo

- How many users are impacted and top bubble categories (Action item)
- Testing and validation of algorithmic bubble identification & classification (load + recent watch list) (Action Item)
- Better understand how users get into bubbles, how users get out of bubbles (if they ever do), and how long bubbles typically last (Action Item)
- Research bubble user demographics, types and behaviors/characteristics of bubble users (Future)
- Adjacent content categories - better understand content categories that are highly related (Future)

Trust & Safety

- Identification of other harmful bubble topics, including: sexualized content, illegal & regulated, hate (Action Item)
- Tools and metrics to monitor platform health from the user consumption perspective, in addition to the traditional content creation focus (Future)
- Use bubble detection technology for early detection of harmful trends and misinformation (Action Item)
- Monitor on and off platform trends to give product and algo team bubble topics to search for (Action Item)

Brief Report on the Filter Bubble in Social Network Platforms