Auditing a Feral Recommendation System

Johan Michalove
michajoh@cs.uw.edu

Matthew Rockett
rockettm@cs.uw.edu

1 Introduction

Contemporary Recommendation Systems (RSs) are often evaluated on their efficacy with respect to performance metrics such as Mean Squared Error and Click Through Rate [1]. While these metrics are easier to track and improve, performance improvement along these metrics does not explicitly account for other factors a user or group of users may care more about, such as diversity of content or its moderation criteria. One result of improving an RS with respect to the traditional criteria is that "filter bubbles" and "echo chambers" can be formed by the RS pigeonholing users into specific clusters of content. We deem a recommendation system "feral" when it is an actor in a complex social milieu yet is goverened by opaque metrics which do not model the interests of their users beyond engagement-centered objectives.

In the last year, platforms have been met with a surge of criticism and calls for oversight regarding their often opaque recommendation and content moderation practices. Although specific criticism of content moderation (or lack thereof) have been levied against platforms, there are few cases of third party action which systematically characterizes the behavior of the recommendation system as an actor in a wider milieu. The measure of a RS's behavior at macroscopic scales is difficult to measure, and the detection of them in a principled approach is a relatively recent academic area.

With the use of deep neural networks for recommendation [2], it is becoming harder for both users and creators of recommendation systems to understand how and why content is being recommended [3]. We look to model users' interactions with YouTube and try to gain an understanding of the recommendation system from the perspective of a community of users, as modeled by a range of simplified engagement patterns using random walks. We hope this will empower users and third parties to explore ways of auditing recommendation systems and keeping these recommendation systems accountable.

With this project we contribute the following:

1. A definition of recommendation networks, building off prior work [4] that takes into account user level interactions with the recommendation system over time.
2. Create and apply a data collection tool to collect instances of a recommendation system for any subject keywords from YouTube.
3. Propose metrics to measure controversy, popularity, and content diversity within the recommendation network framework.
4. Establish hypotheses about recommendation system behavior which advance the notion that algorithmic audits are an essential tool for understanding opaque industrial Artificial Intelligence.

2 Background and Related Work

Auditing methodologies have been proposed as a possible avenue for identifying discrimination on online platforms and as a way of increasing transparency of the role of algorithms in question [5]. One reason these types of audits are called for is that large-scale systems, many of which use machine learning techniques, are regarded to be opaque by virtue of the scale and structure of the

algorithms which are used [3]. This line of thinking has been extended to recommendation systems in a number of studies which attempt to quantify first-order effects of the recommendation system, such as filter-bubbles [6], over specialization and concentration bias [7] and serendipity [8] among others.

A new line of inquiry called Machine Behavior aims to regard algorithms within the framework of sociotechnical, or "hybrid", systems. This focuses on how humans shape machine behavior, how machines shape human behavior and co-behavior of human and machine [9]. One particular system which is pertinent to questions all of these behaviors is the YouTube recommendation system as it is one of the largest Machine Learning based industrial recommendation systems active on the web [2]. Several projects have attempted to shed light on aggregate trends of YouTube's recommendation systems through daily scrapes of recommendations for popular subjects [10] and random walks from a list of popular channels [11]. Other studies have looked at treating recommendation systems as a "black box" and analyzing them purely based on their behavior with (sometimes "mock") users [12, 4].

One particular paper looks to combine the fields of third party algorithmic study and network analysis [4]. In this paper the authors treat a RS as queryable black box. From this black box they build a partial "view" of the system building a "Recommendation Network" – a graph where nodes are items and directed edges are recommendations from the source item to the recommended item – from data collected via many queries to the RS. This reformulation of interactions with recommendation systems allows the authors to apply the rich field of network analysis to analyze the recommendation network [13]. The authors specifically look for parallels between complex network and prior RS analysis for metrics of diversity and information segregation. The authors used Netflix, Google Play, and IMDb to test their framework, allowing for a cross-sectional study between the platforms. In the next sections we will formalize recommendation networks, define our metrics, and describe how we applied this analysis to YouTube's RS.

3 Formulation of the Problem

In this section, we describe the setup of our network-based framework for evaluating recommendation systems. We then discuss the user models used for interacting with the RS with our framework to enable a large scale cross-sectional study. Finally, we describe the auditing methodology and data set we use with this framework.

3.1 Recommendation Networks

In this work, we extend previous works [4] defining a Recommendation Network (RN). A RN is defined as a directed graph $G = (V, E)$ with no self-loops, where $v \in V$ is a video with associated metadata (likes, dislikes, views, etc.), and a directed edge $(u, v) \in E$ is a recorded instance of video u recommending video v. We also incorporate recommendation order into our graph. This allows us to model interactions with the RS that reveal how recommendation order effects user's outcomes.

In the case of YouTube the structure of the RN is created by starting with a search term and following recommendations to a specified depth (more about this in subsection 5.2 data collection section). These video from the search page are our "root nodes."

Unlike the previous definitions, we do not extract from our RN a global view of the RS, but rather an instance of a specific interaction with the YouTube RS. This requires us to look beyond metrics like PageRank and other teleportation surfers to analyze the RN. In the next section we describe our random walker user model.

3.2 User interaction models

In order to analyze the underlying structure of the graphs, we use a random walker model to traverse the graph and collect data on instances of random walks. We define a random walk as an ordered list of nodes $v_1, v_2, v_3, ..., v_k$, where v_1 is a root video and v_k is a leaf node. The transition function from one node to another $Next(v)$ is at the heart of our analysis. We look at a multitude of transition functions, hinging on the ordering of recommendations.

Full random walk In order to collect baseline metrics, we have a random walker that picks an arbitrary video in the list of 20 recommendations as the next node.

Top-k random walk Because the recommendations are ordered by the RS, we use this to gain insights on the RS. A Top-k random walker is similar to the random walker, except it only chooses videos from the top k ordered recommendation set.

4 Metrics

In this section, we describe the metrics used to evaluate the general properties of the YouTube RS across the various walk strategies proposed. These metrics draw both upon inferences which can be made using the mere structure of the network, and also upon inferences which depend on semantic information communicated by the video's metadata, such as likes, dislikes, video category and keywords.

The purposes served by these metrics are two-fold: first, to characterize general trends about the RS which can be monitored over time. The general behavior of RSs change with performance updates, model changes, and policy modifications. Quantitative metrics concerning "improvements" can be tracked using metrics such as those proposed. The secondary goal of these metrics is to understand the nature of possible harms which may occur because of the RS's hybrid-system dynamics: namely, to understand the nature of user-recommendation system co-behavior and the harms which may emerge in the absence of systemic mitigations. Lacking access to real user data, the walker model serves as a simplified substitute.

4.1 Controversy Score

For a video v, with that has l_v likes and d_v dislikes, the controversy score $C(v) \geq 0$ is defined as:

$$C(v) = \frac{l_v + d_v}{\max(1, |l_v - d_v|)}$$

Lower controversy scores correspond to lower user engagement (small l_v and d_v). Videos with low like/dislike engagement will have lower controversy scores regardless of the difference between $|l_v - d_v|$ by nature of being small in magnitude. Videos with higher engagement (larger l_v and d_v) will have high controversy scores when $l_v \approx l_d$.

This metric is used by Reddit.com for ranking comments or posts by controversiality, where the value is 1 if there is full agreement among the votes and "total votes" if there is an equal split between likes and dislikes (the most ambiguous possible situation). This is a reasonable proxy for controversiality because highly controversial videos will have high engagement and simultaneously high polarization in user sentiment, as expressed via a large number of likes and dislikes [14]. We decided to use a metadata-based controversy score, but network based metrics are also available [15, 16].

4.2 Popularity

The popularity of a video can be quantified both with respect to the quantity of its consumption on the platform or with how it's recommended by the RS. We thus propose two distinct metrics for evaluating popularity: number of views and the in degree of videos in the recommendation network.

Number of Views A video v has σ_v views. This serves a proxy for popularity with users, as it indicates the number of instances a user has accessed the video. YouTube only counts views it deems are engagements from "actual humans and not computer programs" – to maintain the quality of this metric, the platform makes its process for counting views intentionally opaque [17].

Recommendation Prevalence A video $v \in V$ in the recommendation network $R = (V, E)$. Let the recommending set $I(v) = \{u \in V | (u, v) \in E\}$, that is, the set of videos in the recommendation network that have presented v as a recommendation. Thus, the number of recommendations made to a video throughout the recommendation network is equivalent to the in degree of v, or $|I(v)|$.

This metric is meant to capture the prevalence of recommendations made to a video throughout the full recommendation network. In one sense, this captures which videos are most popular for the RS to recommend in one or more browsing sessions.

4.3 Content Diversity

Critical questions confronting the study of RSs include how they may lead to "filter bubbles", and what are dynamics that lead to filter bubble formation. One proxy for "being in a filter bubble" is a reduction in diverse content. The first question to address when looking at diversity of content is what similarity metric is used to evaluate the data set. This usually includes salient features of the recommended items which, in the case of the YouTube RS, we've selected two:

1. *Video category:* The platform requires that each video must be assigned one of 15 categories, which describe broad "topics" of video content, such as "Entertainment", "Education", "Science & Technology", "Music", "Nonprofits & Activism", etc. The primary advantages of using the YouTube video category is that there is a label associated with every video and they're standardized across the platform. The primary disadvantages of grouping videos by category are that the labels capture very little detail about the video and, because they are set by users, criteria for what videos belong to a category will vary.
2. *Topic modeling:* Users are free to associate tags with their videos. These generally contain semantic detail related to the content of the associated video. To better understand the content of visited videos, we extracted 45 distinct topic from the dataset using topic modeling. Our procedure for generating topics was to generate TF-IDF features from the 1500 most common features and then apply Non-Negative Matrix Factorization (NMF) [18]. This method was remarkably effective for topic extraction, and yielded better performance on our small (500k documents) dataset than Latent Dirichlet Allocation (LDA) [19].

   By looking at the top keywords associated with each topic, we were able to infer specific semantic groupings, such as "Late Night TV", "Marvel Comic Universe", "True Crime", "Space", "Amazing Compilations" and "Fortnite". Using NMF for topic modeling allows us to rank topics for each video in accordance with likelihood, allowing us to either classify a video as belonging to a specific topics or an ensemble of topics.

In order to study how the diversity of recommended content changes within and between RNs, we propose the following metrics:

4.3.1 Likelihood of recommendation sameness

We can leverage the structure of the RN to look at diversity among recommendations. Using the video category we can look at how the RS recommends videos at varying depths or in various random walks. For a video v we can categorize the category sameness of the recommendations as:

$$S(v) = \frac{|\{(v,u) \in E \,|\, Sim(v,u) = 1\}|}{|\{(v,u) \in E\}|}$$

Where $Sim(v,u)$ is one of the metrics described in the previous section. For video category sameness, $Sim(u,v)$ is 1 if the category of u is the same as the category of v. For topic modeling, we chose to use the two similarity metrics. The first similarity metric "Topic 1", $Sim(u,v)$ is 1 if the highest ranked topic agrees between u and v. The second similarity metric "Topic 3" $Sim(u,v)$ is 1 if the topics agree for at least one of the three highest valued topic scores for the videos. Because topic modeling is only an approximation of the item's topic, and an item can have multiple topics, the "Topic 3" similarity metric attempts to capture similarity in the case that at least one significant topic is shared between two items. For our purposes a binary similarity metric is sufficient, but this could be extended to continuous similarity metrics with a thresholding term.

4.3.2 Assortativity coefficient

Diversity and information segregation in an RS have been operationalized to identify whether a RS produces a reduction of diversity in the items made available. Researchers have used the assortative mixing metric to evaluate the diversity of an RN [4]. Assortative mixing can show trends in how members of one group interact with members of other groups others. We can use assortative mixing

to study how videos in the recommendation network relate to each other by category. The assortativity of a group of videos in the network can be visualized using an *assortative mixing matrix*.

We can also use this strategy to quantify diversity in the system as a whole: The assortativity coefficient [20] is a measure of how nodes with categorical attributes mix. A mixing matrix for a node attribute $\mathbf{e}$, where $\mathbf{e}_{ij}$ is the proportion of edges $(u, v) \in E$ such that, u has value i and v has value j out of the total number of edges. The assortativity coefficient $r \in [-1, 1]$ is defined:

$$r = \frac{\text{Tr}(\mathbf{e}) - \|\mathbf{e}^2\|}{1 - \|\mathbf{e}^2\|}$$

Where $\|\mathbf{e}^2\|$ is the sum of all elements in $\mathbf{e}^2$. When $\mathbf{e}$ is perfectly assortative $\text{Tr}(\mathbf{e}) = 1$, meaning there is no diversity among edges, for every edge $(u, v) \in E$, u and v have the same attribute value. When $r = 0$ When $\mathbf{e}$ is disassortative, $\text{Tr}(\mathbf{e}) = 0$ and $r \leq 0$, implying more diverse mixing, although this doesn't show up in practice for our analysis of the YouTube RS.

Because the assortativity coefficient can be computed for any categorical attribute, we can compute the coefficient for any categorical variable to measure diversity.

5 Data Collection

In this section we detail our data collection process. We first discuss our method for traversing YouTube recommendations, and then our method for extracting structured information from the collected information. We also discuss our ways of composing multiple recommendation networks to aid a wider corss-sectional analysis between subjects.

5.1 Search query selection

Each random walk starts with a video suggested in response to a search query made to YouTube. We selected these search terms using the top 25 related query lists provided by Google Trends for YouTube queries. Thus, we would begin with a single query subject, such as "sports" or "unboxing", then collect the top 10 related queries, and generate exhaustive random walks starting from videos proposed by searching the query.

We chose to do a single, large data harvest and run our random walks in this instead of doing the random walks directly on the platform for two reasons:

1. Efficiency: running millions of walk strategies on YouTube over the network would take an unduly long time. We estimate 16 days of continuous crawling, if performed on a single thread.
2. Consistency: YouTube's algorithm can change from day to day without any external indication of the change. It's difficult to compare runs over multiple days because the way in which the RS presents recommendations could have changed. By collecting a "locked-in" structure of the recommendation network, we can reduce variance introduced by fluctuations in the algorithm and platform activity, such as trending videos. A useful future study would compare a comprehensive random walk experiment to our large-harvest approach.

5.2 Data Harvesting

We collected the relevant data by first gathering popular search queries for a variety of different subjects. We then create an exhaustive recommendation network for this query by following the top 6 recommended videos for the top 6 videos given when searching with the query. We collect data for each video by querying YouTube for a video page, then downloading and parsing the HTML content. The search is done depth first, creating a "surf graph" from the initial recommendations. Each surf graph constitutes an exhaustive trial for a subject, and we repeat multiple trials with different search queries in order to create a more holistic representation of the recommendation network for a subject. After the surf, we collect the data of the many videos that the RS recommended but which we did not visit. Thus the surf graph for each query term has each video completely catalogued.

Once collected, we collect the surf graphs into a data structure amenable to network analysis (`networkx`). We can then collect the metrics defined in section 4, averaging over multiple trials, each

Table 1: Assortativity coefficients for YouTube Categories and Topics.

Subject	**Global**	2020 President	Abduction	Abortion	Avengers	Bikini	Climate Change	Flat Earth	Ford Truck	Hiking
Category	0.50	0.46	0.41	0.40	0.40	0.48	0.42	0.41	0.51	0.38
Topic 1	0.34	0.31	0.32	0.28	0.45	0.34	0.29	0.30	0.30	0.31
Topic 3	0.62	0.60	0.59	0.58	0.67	0.61	0.58	0.59	0.59	0.59
Subject	History of War	Homeless	Reptiles	SpaceX	Sports	Tesla	Unboxing	Unemployment	Us Movie	Volcano
Category	0.44	0.47	0.48	***0.29***	**0.62**	0.47	0.58	0.41	0.40	0.46
Topic 1	0.28	0.34	**0.47**	0.29	0.41	0.32	0.36	***0.25***	0.33	0.33
Topic 3	0.57	0.61	**0.68**	0.57	0.65	0.60	0.63	***0.55***	0.62	0.61

The first entry is the global value (calculated over all RNs). The others are calculated by merging the RN for the 10 search terms and calculating the assortativity. The highest assortativity coefficient for YouTube Category and Topic are **bold**, the lowest are ***bold and italicized***.

of which is collected on an individual surf graph. Across the 19 subjects, we collected the data of 5,064,291 videos, recorded 12,884,067 edges, collected 190 surf graphs that totaled nearly 2.5 GB and required 48 hours to harvest.

We plan to make our system open source so that others can replicate our experiments and collect data with other surfer models, collections of query terms, etc. We hope this will allow third parties to not only study the network for individual subjects at a single point in time, but also to do longitudinal studies of how the recommendation system changes.

6 Results

In the following section we show our evaluation of YouTube's RS using our RN metrics and methodology. The fist section will talk about overview metrics across the entire network. We then apply the same metrics over multiple user interaction models to understand how interaction strategies with the RS change the attributes of which videos are recommended.

6.1 An overview

We begin by looking at an average of the metrics over each surf graph collected across the explored depths 1-6 to get an overview of the mean and variability of the metric. The data as represented in Figure 1 provides us with a reference point when viewing the metrics across interaction models.

We see the error decreasing as one traverses to lower depths. We suspect that this is caused by the RS collapsing uncertainty about optionality with respect to possible recommendations due to the surfer providing additional feedback to the system. We note in Figure 1a that after following two recommendations the controversy score of items at this depth converges to 1.5. We find, in our random walk results, presented in subsection 6.3, that this controversy score has little deviation from this quantity for a majority of the walk strategies attempted. For the subjects we studied, we can conclude that on average the RS does not suggest controversial videos, or, videos which are recommended are not controversial. The ability for the RS ensure a viewer is not recommended controversial content is chiefly accomplished by reducing variance within the first two selected recommended videos.

It is worthwhile noting that our controversy metric judges videos as controversial with respect to the amount of *polarization* as expressed by likes and dislikes on the video. Thus a video can be uncontroversial to the audience which has viewed–perhaps it even has near unanimity in its "likes" expressed by viewers–while being highly controversial to a broader audience.

The mechanism for placing users among videos that she will not find controversial is suggested by Figure 1c. We see that videos that were recommended on the pages of root videos have a significantly higher recommendation prevalence as expressed by their in degree in the surf graph. This is a result of these videos appearing frequently throughout the surf graph. We suspect these videos play an important role in cold-start profiling user preference, which would also contribute to reducing variance in controversy scores at lower depths as recommendations adapt to the uncovered preferences.

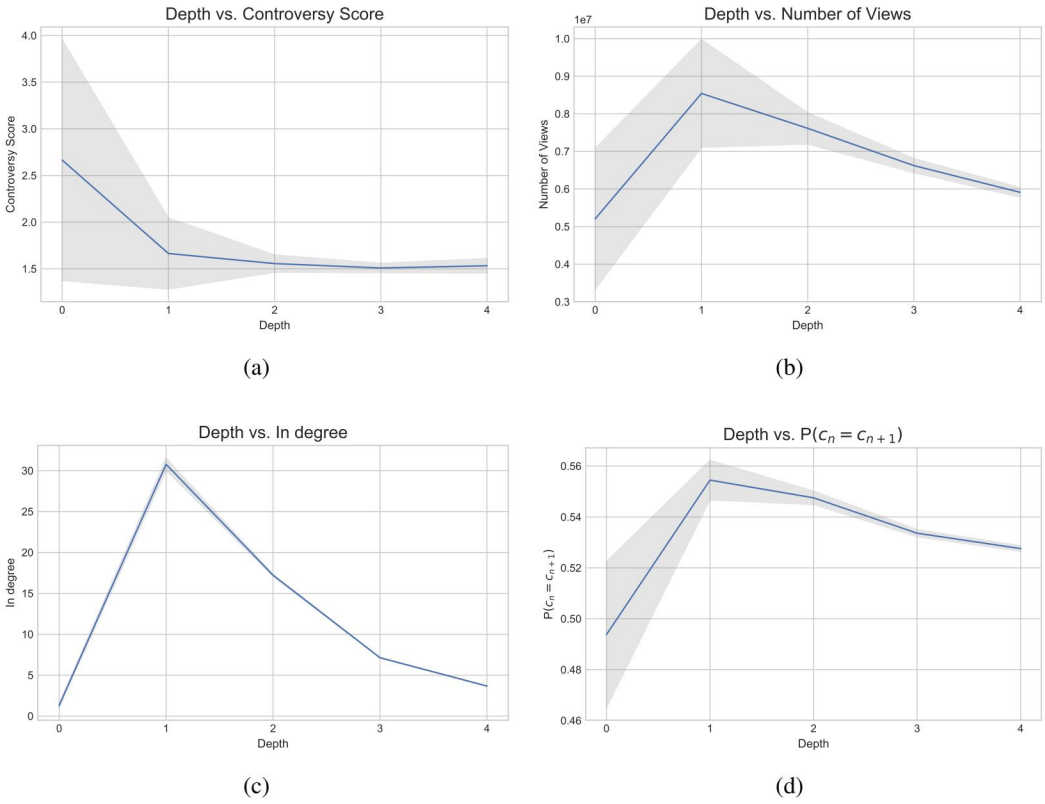


Figure 1: An overview of measurements aggregated over all collected recommendation networks. The shaded regions are the 95 percentile of the data.

Figure 1b further reinforces this hypothesis because there is a spike in the mean and variance of views at depth 1, suggesting that the RS presents videos with higher user popularity as expressed by the number of views. This is an advantageous move for cold-start as these items contain more information about user interaction with the video.

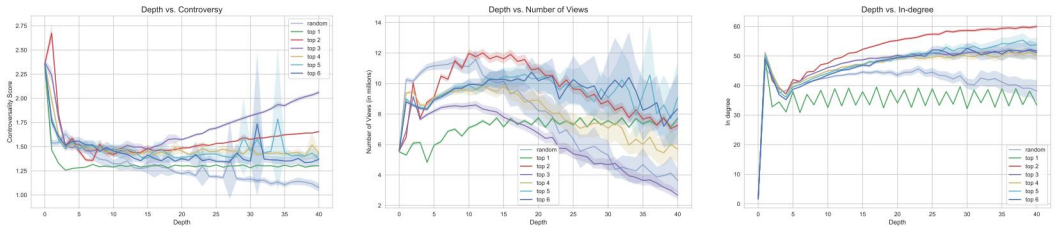
We see in Figure 1d that the likelihood that a video's category changes is approximately 0.5 for all explored depths. We suspect that this is because other latent factors play a much larger role in determining which video gets suggested. As the RS has access to a rich feature set for decision making [2], it's not surprising categories are, on average, an ineffective proxy for determining which videos will be suggested.

6.2 Assortativity coefficient

The assortativity coefficients for the combined "Global" graph and each individual subject graph are given in Table 1. We merged the surf graphs of each query belonging to the subject, and then measured the assortativity coefficient with respect to the similarity measures described in subsubsection 4.3.1 where $\mathbf{e}_{ij} = |\{(u, v) \in E \,|\, Sim_{ij}(u, v) = 1\}|$ where $Sim_{ij}(u, v) = 1$ if $a_u = i$ and $a_v = j$ where a_u and a_v are attributes of videos u and v which have the attributes with the value i and j respectively.

The assortativity coefficient serves as a proxy for diversity in the RN, allowing for a view into which query subjects yield more or less diverse recommendations. We can compare this to the "Global" subject, which captures the assortativity across all recommendations observed in every surf graph we collected.

We observe that the most diverse subjects are "SpaceX" for category and "Unemployment" for both topic similarity scores. The least diverse subjects are "Sports" for category and "Reptiles" for topic similarity scores. Moreover, we notice that the global assortativity coefficient nearly doubles when



(a) (b) (c)

Figure 2: Our measurements with different walk strategies. Depth versus (a) Controversy Score; (b) Number of Views; (c) In-Degree. Shaded area is is 95 percentile of data

going from "Topic 1" to "Topic 3" similarity metrics, suggesting that results from the RS are more likely to be of a similar topic than not.

For third party auditors invested in tracking notions of "fairness" in recommendation systems with respect to the diversity of suggestions made by an active RS, creating such an assortativity table for significant subjects can be useful for tracking the relative diversity between subjects on the platform. If there exists significant discrepancies in assortativity coefficients between subjects, it may be cause for concern that some subjects are more prone to filter bubbles than others.

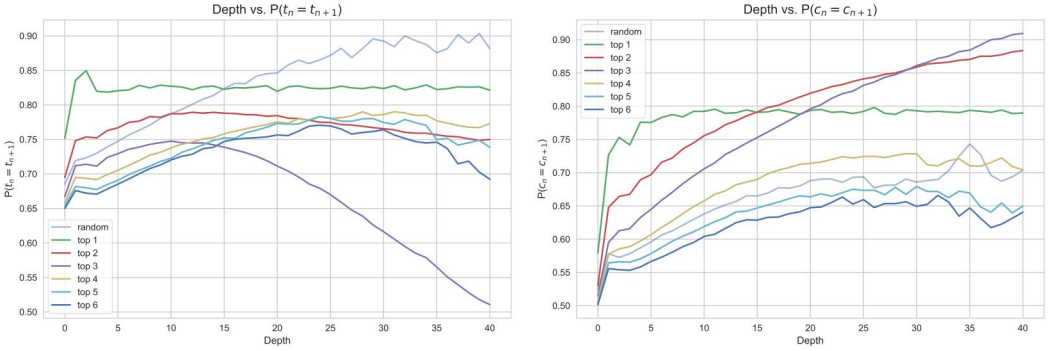
6.3 Walks

For the random walk analysis we simulated 5000 random walks on each surf graph for each walk strategy, amounting to 6,650,000 walks in total. We focused on 7 random walk strategies: (1) Random walk; (2); top-1 walk; (3) top-2 walk; (4) top-3 walk; (5) top-4 walk; (6) top-5 walk; and (7) top-6 walk. Because we only inspected the top 6 recommended videos in data collection, we limited our walks to choosing from these videos in the top-k strategies. Figure 4 shows the relationship between topics and categories for select walk strategies.

In plotting our results, we focused on the metrics up to depth 40. Although there were walks that continued out to higher depths, these walkers were usually stuck in a cycle, and would traverse it until choosing a random recommendation that ended the walk. These "cyclers" accounted for less than 0.1% of the random walk data.

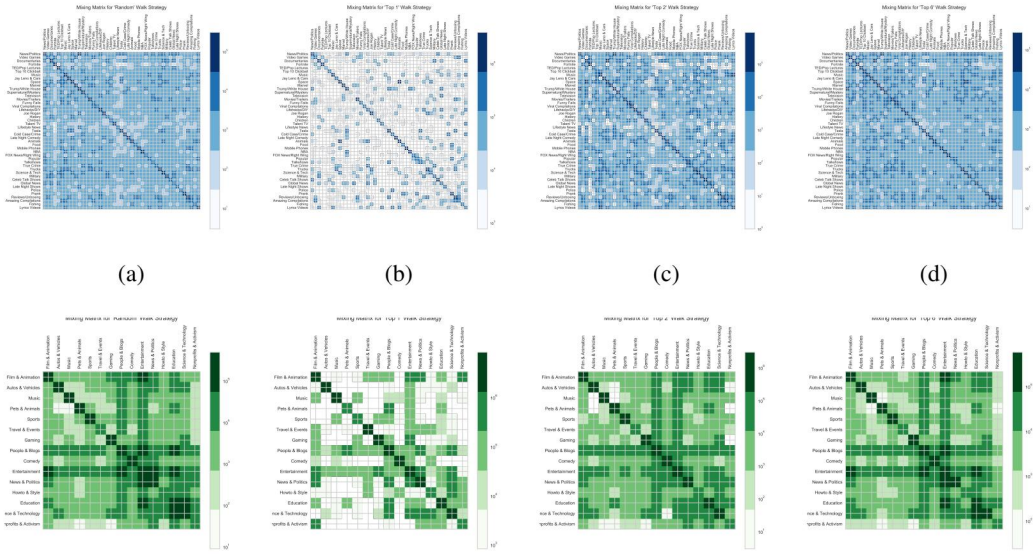
The most significant insight yielded by evaluating the RS across multiple random walkers is that the top 1-3 recommendations yield notably different trends than when the walker uses any of the other walk strategies. In particular, however, the top-3 walk strategy yielded outlier results for all metrics but recommendation prevalence, where top-2 visited the most prevalent videos. When testing walk strategies with respect to the controversy score, top-2 and top-3 were the only strategies that yielded a sustained positive slope, meanwhile random walking was the only which yielded a steady decrease. These results are displayed in Figure 2a. Comparing top-2 and top-3 suggests that the introduction of the third recommended video will lead a user to more controversial and less viewed videos as the user progresses.

One hypothesis for this result is that because the top 3 videos are seen by the user without scrolling, the user is most likely to accept a suggestion from these top 3. Therefore the RS hedges between familiar and novel content in order to ensure that users bored with the current content can "explore" with minimal effort, while the RS "exploits" prior knowledge about popular videos by suggesting likely, familiar videos. We suspect that the first two videos are biased towards the familiar, as indicated by Figure 2c which yields high recommendation prevalence whereas all other strategies are more balanced. Figure 3a reinforces the hypothesis that the third video is a recommendation outlier. In future work, we would like to run the random walks strategies directly on the platform to validate that these trends are not artifacts from our collection methods.



(a) (b)

Figure 3: Our measurements with different walk strategies. Depth versus (a) Likelihood of topic sameness; (b) Likelihood of category sameness. Note that walk strategy top-3 suggests that recommendations of diverse topics often takes place within the same video category.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4: Visualization of the mixing matrices for different walk strategies. For any element i, j in the mixing matrix M, the value M_{ij} is the number of edges $(u, v) \in E$ where u video has topic i and v has topic j. (a) is the distribution of edges for the random walker; (b) is the distribution of edges using the top-1 walker (note, the top-1 walker is deterministic, which is why there are few elements with large values); (c) is the distribution for the top-2 walk strategy. (d) is the distribution for the top-6 walk strategy. Note that even with logarithmic scaling, diagonal entries feature most prominently for nearly all categories and topics. *For best view, use a high zoom in PDF viewer.*

7 Conclusion

Although industrial RSs are notably opaque in their behavior and ranking, the network-centric framework we applied to the YouTube recommendation system provides a promising set of tools and metrics for algorithmic auditing. The hypotheses advanced by this paper would have been difficult, if not impossible, to infer from looking at individual cases of recommendations on the platform. This advances the argument that effective scrutiny of online platforms which rely on opaque ranking and recommendation systems require external, third party audits which go beyond critiques of individual cases of content moderation failures, anecdotal evidence, or other first-order observations. Although the sampled data collected in this paper accounts for a miniscule amount of information relative to YouTube's usage, it illuminates how the YouTube RS behaves as an actor in its social milieu. We maintain that until the behavior of the RS is observably acting with motives that indicate it is modeling users interests beyond engagement-centered measures, it remains feral.

7.1 Limitations

While this project allowed us to collect a large amount of data on the YouTube RS, there are a few limitations we would like to acknowledge.

Authentic walks The random walking strategies we tested were simplified strategies meant to model user activity. A future paper would ideally use real user browsing data in lieu of random walks. One ought to answer the question: to what extent do random walker models need to be faithful to browsing activity to capture the underlying hybrid-system interactions? Our data harvesting method, while efficient, may also impose certain limitations on the authenticity of recommendations due to the RS's complex feedback mechanisms.

Subject selection We only had a limited number of subjects that we chose based on Google Trends data. While this allowed us to gain insight on a range of content that are prevalent on the platform, this is a limitation biased by what we thought represented a diverse selection of subjects. The selection of which subjects third party auditors study may in large part depend on the interests of the investigating party. In the future, one could take a more systematic approach to exploring the topic space or work directly with communities of direct and indirect stakeholders [21].

7.2 Future Work

Longitudinal study We want to not only evaluate a system at any given time, but see how it evolves over time. We hope to use and extend the tool to allow researchers to study recommendation systems not as a static entity, but as a dynamic system with mutual feedback with users and other machines.

Incorporate rich user model Our data collection was agnostic to how users interact with the system. While, our data collection tool allowed us to gather a large amount of recommendations from a pool of topics, it does not emulate user's experience with YouTube, and how their interactions with the RS effect the RS. In the future we would like to incorporate data collection that allows us to collect a dataset with richer interaction from the data collection tool and the system.

7.3 Member Contributions

We feel that we equally contributed both time and material contributions to the paper. We spent most development and writing hours of this project working on this project together.

1. Johan Michalove: defining metrics, paper writing, data analysis framework, topic modeling, problem formulation
2. Matthew "the rocket" Rockett: data collection framework, data collection, plotting, random walking analysis, problem formulation

References

[1] Guy Shani and Asela Gunawardana. Evaluating recommender systems. Technical report, November 2009.

[2] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, New York, NY, USA, 2016.

[3] Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, 2016.

[4] Abhisek Dash, Animesh Mukherjee, and Saptarshi Ghosh. A network-centric framework for auditing recommendation systems. *CoRR*, abs/1902.02710, 2019.

[5] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cédric Langbort. Auditing algorithms : Research methods for detecting discrimination on internet platforms. 2014.

[6] Tien T. Nguyen, Pik-Mai Hui, Franklin Harper, Loren Terveen, and Joseph Konstan. Exploring the filter bubble: The effect of using recommender systems on content diversity. pages 677–686, 04 2014.

[7] Panagiotis Adamopoulos and Alexander Tuzhilin. On over-specialization and concentration bias of recommendations: Probabilistic neighborhood selection in collaborative filtering systems. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 153–160. ACM, 2014.

[8] Tomoko Murakami, Koichiro Mori, and Ryohei Orihara. Metrics for evaluating the serendipity of recommendation lists. In Ken Satoh, Akihiro Inokuchi, Katashi Nagao, and Takahiro Kawamura, editors, *New Frontiers in Artificial Intelligence*, pages 40–46, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[9] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. Machine behaviour. *Nature*, 568(7753):477–486, 2019.

[10] Guillaume Chaslot. Youtube's most recommended videos.

[11] Aaron Smith, Skye Toor, and Patrick van Kessel. Many turn to youtube for children's content, news, how-to lessons, Dec 2018.

[12] Namhee Lee, Jason J. Jung, Ali Selamat, and Dosam Hwang. Black-box testing of practical movie recommendation systems: A comparative study. *Comput. Sci. Inf. Syst.*, 11:241–249, 2014.

[13] M.E.J. Newman. The structure and function of complex networks. *Computer Physics Communications*, 147:40–45, 03 2003.

[14] reddit. `https://github.com/reddit-archive/reddit/blob/master/r2/r2/lib/db/_sorts.pyx`, 2015.

[15] Shiri Dori-Hacohen and James Allan. Detecting controversy on the web. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1845–1848. ACM, 2013.

[16] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):3, 2018.

[17] How video views are counted - youtube help.

[18] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[19] David M Blei and John D Lafferty. Topic models. In *Text Mining*, pages 101–124. Chapman and Hall/CRC, 2009.

[20] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, Feb 2003.

[21] Batya Friedman, Peter H Kahn, and Alan Borning. Value sensitive design and information systems. *The handbook of information and computer ethics*, pages 69–101, 2008.