Auditing a Feral Recommendation System

CSE 547 - Spring 2019
Johan Michalove and Matthew Rockett

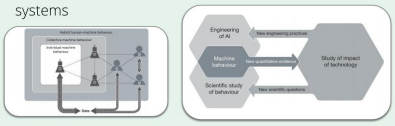
PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Motivation

- **Social effects** of recommendation systems (RSs) are not well understood
- Recommendation systems are usually optimized for **engagement** using "traditional" metrics such as:
  - Viewing time
  - Click Through Rate
  - Mean Squared Error
- Improving an RS with respect to traditional metrics can lead to formation of "**filter bubbles**" and "**echo chambers**" by the RS pigeonholing users into specific clusters of content.
- **RQ: Can we empower users and third parties to explore ways of auditing recommendation systems that holds the RS accountable as an actor in a social milieu?**

Hybrid Human-AI Recommendation Systems

- **Field of Machine Behavior** aims to regard algorithms within the framework of sociotechnical, or "hybrid" systems



- "**Feral" recommendation system:** an actor in a complex social milieu that is governed by opaque metrics which do not model the interests of the users beyond engagement-centered objectives.

Key insight: lack of sociality. Can a RS be your friend?

Recommendation Networks: an analysis framework

The **Recommendation Network** (RN) is a representation of a set of recommendations made by the RS under test.

Problem Formulation: RN and Random Walk

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 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.

Allows for the use of **complex network analysis** to understand dynamics of RS.
Random walker model is a simplified **user interaction model** to study effects of browsing patterns on recommendations.

Metrics

Controversy Score:

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

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

Popularity:

- Number of views
- Recommendation Prevalence: In-degree of item *v* in RN

Recommendation Diversity

Assortativity Coefficient

A mixing matrix for a node attribute e, where 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:

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

Content Diversity:

- Category: attribute provided by YouTube
  Eg. "Science & Technology", "Entertainment"
- Topic Modeling: TF-IDF and NMF topics
  - Eg. "Avengers", "Late Night TV", "Amazing Compilations"

Attributes used to compute $\mathbf{P(a_v = a_u)}$ where $\mathbf{a_v}$ and $\mathbf{a_u}$ are attributes of videos v and u. Sim(v,u) used to determine similarity:

For a video v we can categorize the category sameness of the recommendations as:

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

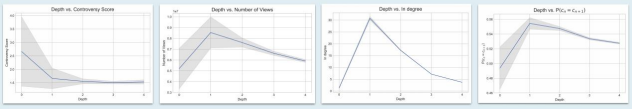
Sim methods: "Category", "Topic 1", "Topic 3"

Walk Strategies:

Random: picks an arbitrary video in the from the first 20 recommendations

Top-K: picks a video from the first k ranked recommendations (k = 1 ... 6)

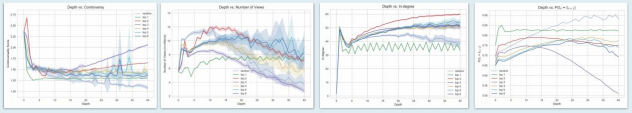
Results: An introduction

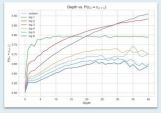


Key insights:

- Uncertainty collapses as the user reaches a lower depth.
- On average the RS does not suggest controversial videos, or, videos which are recommended are not controversial.
- The RS steers users towards highly viewed videos first, before leading them to less popular content.

Key Results





- Top-3 surfer trends:
  - Top-2 and Top-3 walkers, progressively lead to **more controversial** content*.
  - Top-2 walker at shallower depths lead to on average higher viewed videos, while Top-3 leads to **less viewed videos**.
  - Top-2 walker has a much higher prevalence in our surf graph, the RS proposed as a recommendation more often.
  - While the Top-2 and Top-3 surfer consistently get recommended content of the same category, the Top-3 surfer sees more diverse "topics".

Why? Third video plays a special role for recs -- familiar vs. novel

* by our definition of controversial

Data Collection

- We analyze **YouTube** as it is a very popular industrial recommendation system, and has been under intense scrutiny for it's possible negative effects recently.
- **Surf graph:** depth first search for a subject (consisting of 10 search terms). Graph is created by taking top 6 videos from a search query and exhaustively DFS to depath 6.
- We selected 19 subjects, for a total of 190 surf graphs.
- In total collected data on 5,064,291 videos, capturing 12,884,067 recommendations on YouTube

Diversity: Assortative Mixing

Assortative Mixing Matrix describes recommendation frequency between item attributes.

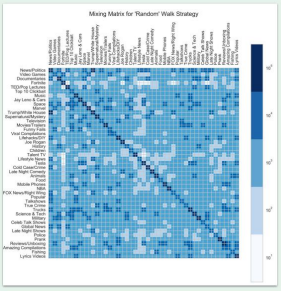


Table 1: Assortativity coefficients for YouTube Categories and Topics.

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

Conclusions and Takeaways

- **Conclusions:** Large scale algorithmic auditing tools are one of several frameworks for effective scrutiny of large industrial RSs.
- **Need for third party audits** which go beyond critiques of individual cases of content moderation failure, anecdotal evidence, or other first-order observation.
- **Limitations:** Use more directed **subject selection** to understand different topics.
- **Future work:**
  - Incorporate real **user data** for creating walk strategies, and running them **on** the platform.
  - Open source so other interested third parties and researchers can leverage our tool.