

Mini-Report: Exploring Filter Bubbles in YouTube Recommendations

Jader Silva

Introduction

This mini-project explores the dynamics of video retrieval and recommendation using three neutral seed queries: *healthy cooking*, *beginner guitar*, and *stretching exercises*. The purpose is to investigate how content clusters form around different seeds, to quantify diversity, and to assess overlap between themes. While simplified, the study illustrates core analytical skills in text processing, similarity analysis, and visualization of recommendation structures, all of which are transferable to larger-scale media ecosystem studies.

Method

We implemented a lightweight pipeline to explore **YouTube recommendation bubbles** across three thematic seeds: “*beginner guitar*”, “*healthy cooking*”, and “*stretching exercises*”.

The procedure was:

1. **Data collection** – Using the YouTube Data API v3, we queried each seed and followed the first layer of recommended videos. Results were stored in structured CSV files (data/raw/ → data/clean/ → data/processed/).
2. **Preprocessing** – Video metadata (title + description) was concatenated into a single text field. Duplicates and empty fields were removed.
3. **Vectorization & similarity** – We applied TF-IDF over the text fields and computed pairwise cosine similarity. This yielded a similarity graph between videos.
4. **Aggregation** – Instead of rendering raw network visualizations (dense and hard to interpret), we extracted **aggregate metrics** (Jaccard overlap, diversity/entropy, largest connected components) to summarize cluster behavior.
5. **Visualization** – Metrics were represented as **heatmaps, bar charts, and line plots**, emphasizing interpretability over raw network density.

Visualizations

- **Jaccard similarity heatmap:** pairwise overlap between seed clusters, showing how much the video sets intersect.
- **Diversity by seed (entropy measure):** normalized entropy indicating how uniformly each seed's recommendations are distributed.
- **Largest Connected Component (LCC) size:** bar chart comparing the scale of the main cluster within each seed's graph.
- **Entropy vs. random walk step:** line plot tracking how recommendation diversity evolves as the user navigates outward from the initial seed.

These plots replaced the earlier exploratory *networkx* diagrams, providing clearer, cluster-level insights.

Results

1. Entropy vs Step

Entropy curves show *stretching exercises* accumulating the highest diversity. *Healthy cooking* starts narrow at early steps but rises sharply by step 3, surpassing *beginner guitar*. This reflects that the step-wise curve captures the path of sampled recommendations, whereas the seed-level diversity metric summarizes the full distribution over clusters.

2. Jaccard Similarity Between Seeds

Video-level Jaccard similarity between seeds was 0, confirming that no videos overlapped across queries. However, channel-level overlap was observed between *healthy cooking* and *stretching exercises* (e.g., recurring creators like Lilly Sabri and Massy Arias), suggesting cross-domain content strategies by certain channels.

3. Diversity per Seed

Normalized entropy scores reinforced the entropy curve findings:

- *Beginner guitar* → most balanced topical spread.
- *Stretching exercises* → moderate diversity.
- *Healthy cooking* → minimal diversity, with videos concentrating on similar phrasing and themes.

4. Largest Connected Component (LCC)

The size of the largest connected component was computed for each seed's recommendation

graph. All seeds produced a single dominant cluster containing the majority of videos, consistent with the formation of tightly bounded recommendation spaces.

Limitations

This study is constrained by its scale: only 3 queries, ~450 videos total, and a reliance on textual metadata rather than full video transcripts or engagement signals. The Jaccard analysis, while informative, captures only surface-level overlap (IDs and channels), not thematic subtleties. The clustering method (KMeans on TF-IDF) provides a coarse partition, sufficient for a toy project but not for production-level recommendation audits.

Next Steps

Future iterations should:

- Expand seeds to politically sensitive or polarized queries, to test recommendation pathways in high-risk contexts.
- Incorporate temporal dynamics (video publication dates, trending status).
- Enrich the feature set with engagement metrics and full transcripts.
- Scale diversity and overlap metrics to larger graphs, enabling network analysis of recommendation bias.