# Recommendation Report for YouTube Suggestions to Decrease Bias and Centralization.

Prepared For:

Mr. Brock Baines

COMM-6019 Instructor

Fanshawe College

1001 Fanshawe College Blvd, London, ON N5Y 5R6

Prepared By:

Riley J. Huston

Fanshawe College

1001 Fanshawe College Blvd, London, ON N5Y 5R6

August 2, 2024

# Table of Contents

# Table of Figures

# Letter of Transmittal

August 2, 2024

Mr. Brock Baines - COMM-6019 Instructor
Fanshawe College
1001 Fanshawe College Blvd
London, ON N5Y 5R6

Dear Mr. Baines,

Here is the recommendation report you authorized for YouTube suggestions to decrease bias and centralization.

I have found various areas which indicate bias within the YouTube recommendation system, including massive biases towards certain categories and channels. 81.2% of the recommendations were from the same category and 76.7% of recommendations were from the same channel. The data also indicates that 25% of YouTube's recommendations will come from the entertainment category and 22% of all recommendations will come from the top 11 channels.

I would recommend adding one of these recommendations: a cumulative variance variable where the farther a user delves into a category or channel the more, they are recommended out of it. Or a random surfer approach, which is a constant variable which gives all recommendations a chance to be outside of the category or channel.

The information found within this report came from online sources as well as my own data gathering program.

Thank you very much for the opportunity to conduct this research for you. I have learned a lot from this experience and would love another opportunity in the future. I will be looking forward to presenting these findings to you within the next two weeks.

Sincerely,

RHuston

Riley Huston

# Executive Summary

To decrease the bias in YouTube’s recommendation algorithm, this report recommends the implementation of either recommendation variance or a random surfer variable.

There is strong evidence that there is heavy bias in what YouTube recommends to its users. It is shown that recommendations will very often be similar to the current video a user is watching, with 81.2% of recommendations being from the same category and 76.7% of recommendations from the same channel. Data indicates that 25% of YouTube’s recommendations will come from the entertainment category and 22% of all recommendations will come from the top 11 channels.

To decrease this bias, YouTube must implement at least one of the following:

**Introducing Recommendation Variance**

- Introduce a weighted variable which increases the longer a user spends in a given category or channel. This variable will be the chance that a recommended video belongs outside of the current category or channel.

**Applying a Random Surfer Approach**

- Introduce a random surfer variable which is a constant chance that a given recommendation is outside of the current category or channel.

# Introduction

YouTube is a massive online video sharing platform, one that just about anyone will be familiar with. There is reportedly 500 hours of content uploaded to YouTube every minute as of February 2022 (Ceci, 2024), and 14 million daily active users which view more than 1 billion hours of video per day as of Oct 10, 2023 (Elad, 2023). The number of users and content on YouTube is staggering.

However, only a very small percentage of that content is viewed and content that is viewed is centralized to a select few channels. This centralization of content does a disservice to the sheer amount of other unseen and unpromoted content that YouTube holds which can lead to echo chambers of information.

Since YouTube is such a large portion of the video media space online, there is a duty to ensure the content suggested is not centralized and comes from a wide range of topics and viewpoints.

Since YouTube's algorithm is a black box, it is impossible to know exactly what methods are used to recommend videos to users, however we can analyze the content it does recommend giving us an understanding of what YouTube tends to recommend to its users. This report will investigate the YouTube algorithm by simulating watching ~100 000 YouTube videos by continuously watching the first video recommended to a guest user, meaning there is no user data associated with the recommendations. We will perform 1000 video increments of this before switching to a new video outside of the recommendations. This process repeats 100 times to give us the total of ~100 000 videos watched. Data from each YouTube video is collected and analyzed through a Python program using YouTube's Data API V3 and the Python library Plotly. The online source "*Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users*" was also used to corroborate the findings of this study.

# Discussion of Findings

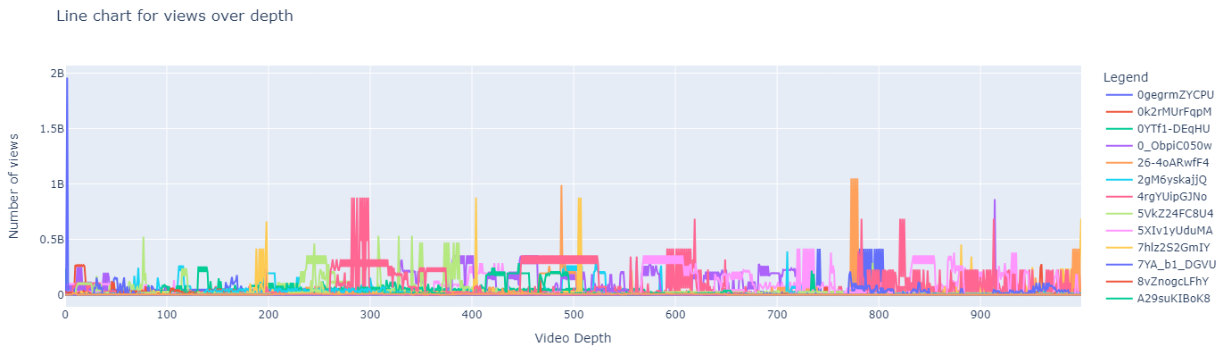
Data from 105 823 total YouTube videos were gathered, the key data points to analyze will be video view count, video category and video channel. Using these points of data, bias can be shown within YouTube's recommendation system. It is important to define some terms that will be used in the discussion of findings. First, a "root video" or simply root refers to the first video in each recommendation path. The video data was gathered by taking the top 100 trending videos at the time of running the simulation, then delving 1000 suggested videos deep for all of them. A root video refers to these trending videos. Secondly, visits refer to the number of times a video was seen in the simulation. This may be confused with views, which refers to the number of views which the video has on YouTube. Finally, depth refers to the number of videos that have been seen in each recommendation path.

It is also important to understand that the YouTube algorithm is in a constant state of updates and changes. These results may differ if the simulation were to run right now. Large societal events such as the Olympics occurring at a similar time as the simulation could also theoretically affect the outcome of the results. The data to be analyzed in this report was gathered from July 11, 2024, to July 13, 2024.

## Number of Unique Videos

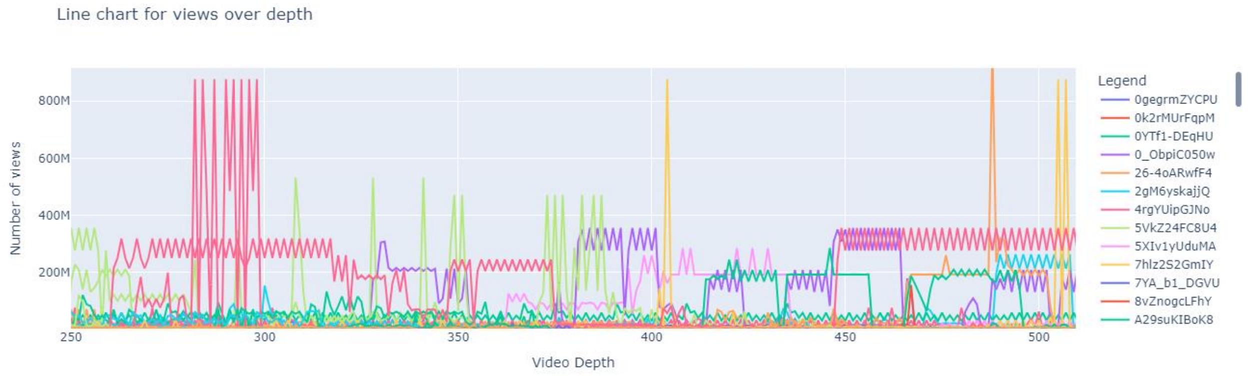
To begin, the total number of videos is 105 823. However, only 19 617 of those videos were unique. This means that only ~81% of the videos are repeats of those unique videos. Already that shows a huge amount of repetition in the recommendations of YouTube videos. This repetition can be seen in this line chart:

*Figure 1 - Line chart for views over depth.*



Each line represents a different root video as you traverse its recommendation path. Each point on a given line is a video corresponding to the current depth and the number of views that video has. Zooming in to a particular section of this plot shows the repetition of video recommendations clearly.

*Figure 2 - Zoomed in section of the line chart*

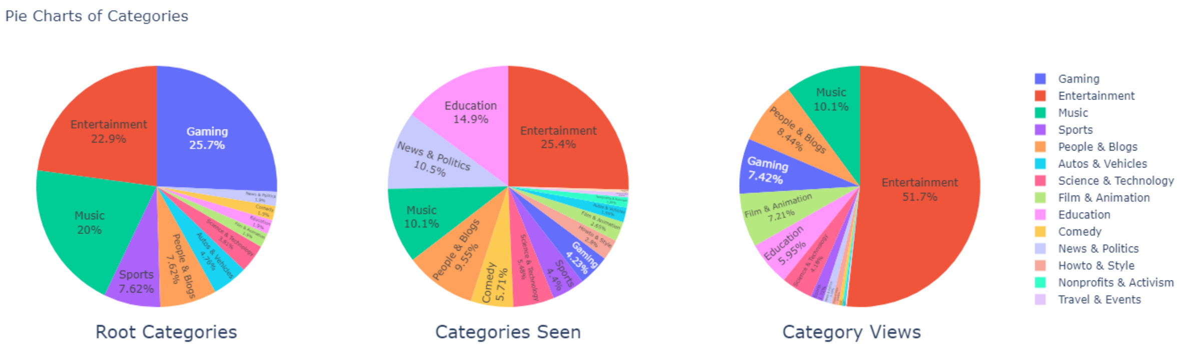


Those prevalent zig zag patterns show that often when a video is recommended, it will recommend back to the previous video. Because of the variance of the recommendation system there is a chance to break out of this loop. However, this establishes a bias towards recommending similar videos to the one the user is currently watching.

## Categories

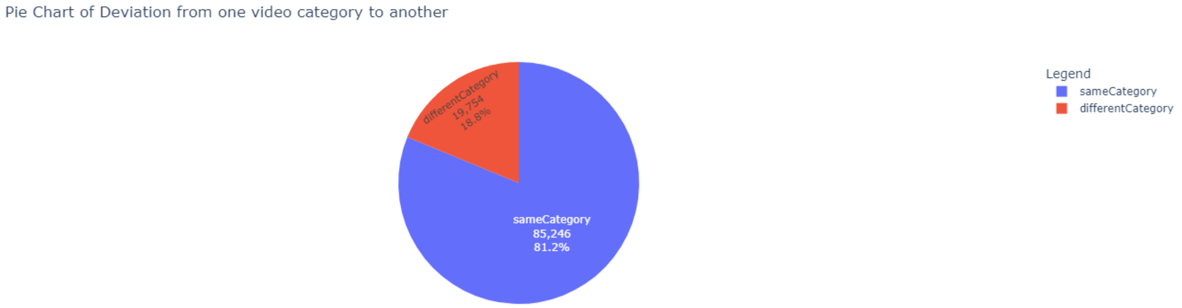
YouTube's Data API provides a video category for each video. The graphing program gathered all the categories and placed them into three pie charts to analyze category ratios. The first pie chart is for root videos, the second is for the categories visited, and finally, the third is for the total category views for our dataset.

*Figure 3 - Pie charts of Categories*



These results show a heavy bias for the entertainment category. Not only are 25% of the 105 823 videos in the entertainment category, 51% of the views also come from that category. 25% of the 100 starting categories were gaming, which dropped off heavily down to only 4.23%, suggesting that YouTube recommended the user out of the gaming category into other categories, which may suggest a negative bias towards certain categories. However, there is no substantial evidence to confirm this observation. The data does confirm that YouTube often will recommend videos within the same category, shown through the following pie chart:

*Figure 4 - Pie Chart of Deviation from one video category to another*

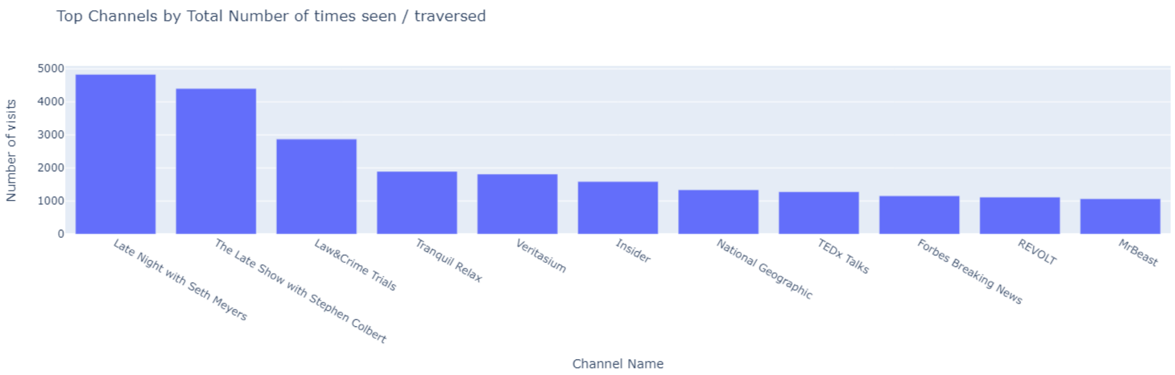


81.2% of all recommendations were within the same category as the current video. This displays an overwhelming bias to feed the user similar content.

## Channels

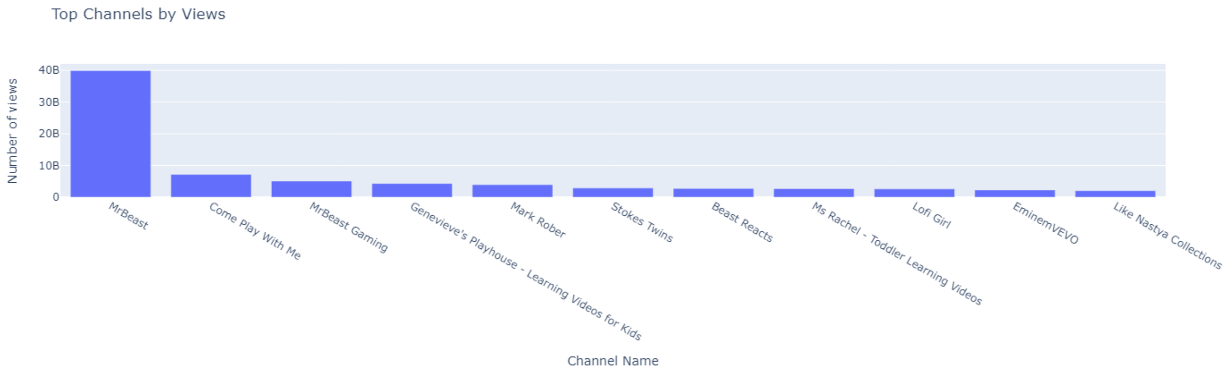
Along with categories, YouTube's Data API provides a channel for each video. There are a total of 3830 unique channels in the dataset. This means that if all channels had the same number of videos, each channel would have been visited ~27 times. The graphing program collected the number of times each channel was seen and placed them in a bar chart ordered from most seen to least. The following are the top channels seen in the dataset.

*Figure 5 - Top Channels by Total Number of Visits*



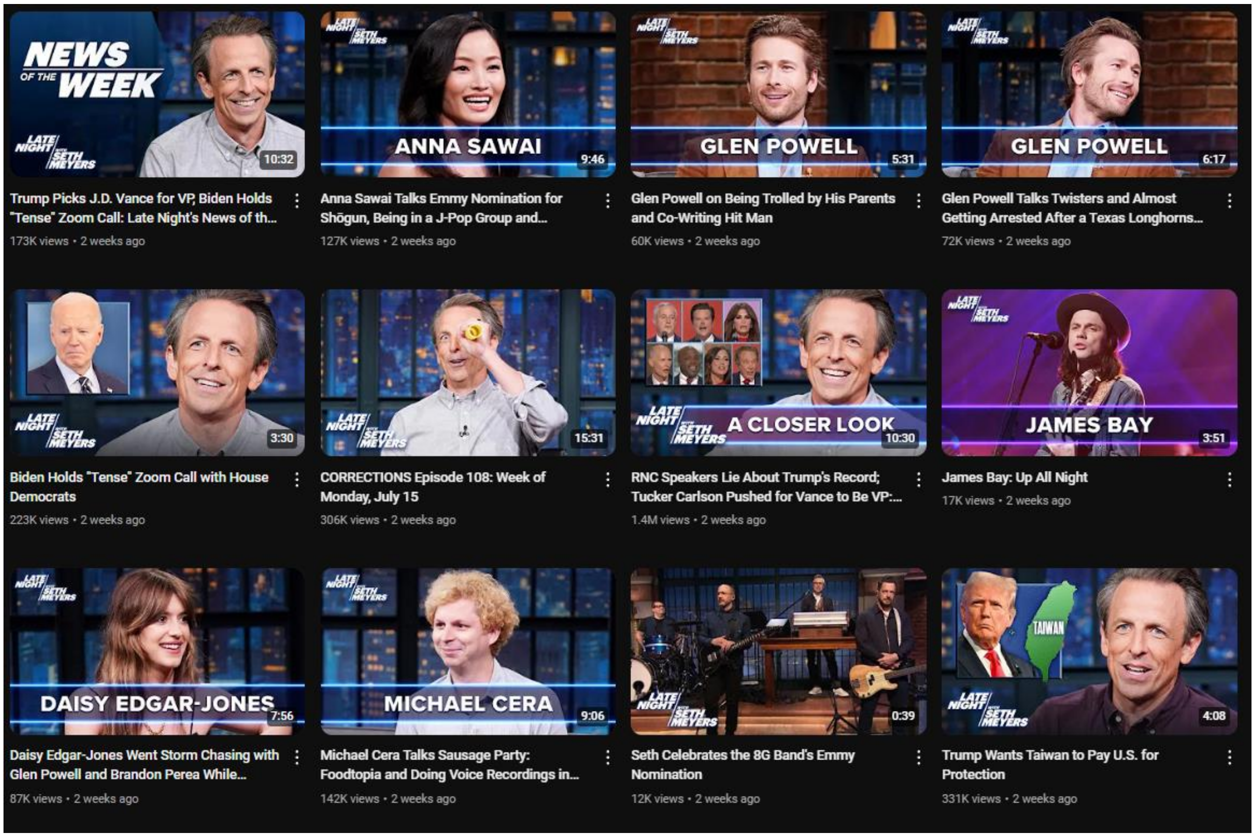
The results from this graph are staggering. This shows that YouTube has a massive bias towards certain channels, as the top two channels were seen 9239 times collectively with a heavy drop off, meaning that they were ~9% of all the videos suggested. If all eleven videos in the chart are totaled, they account for 23 382 of the 105 823 videos, or 22% of all the videos suggested. Since these channels have such a high volume of visits, it would make sense for them to have many views as well. The following graph shows the top channels by their view count.

*Figure 6 - Top Channels by Views*



This shows a massive disparity between the visit count and view count of channels. With the top channel by visits "Late Night with Seth Meyers" not even appearing in this graph, being the 176th by views. This may be influenced by the frequency and volume of video uploads. "MrBeast" does not upload very often, however his views are generally in the hundreds of millions. While "Late Night with Seth Meyers" will upload much more often with multiple videos a day.

*Figure 7 - Screenshot showing that "Late Night with Seth Meyers" upload volume is very high.*



Trump Picks J.D. Vance for VP, Biden Holds "Tense" Zoom Call: Late Night's News of th...
173K views • 2 weeks ago

Anna Sawai Talks Emmy Nomination for Shōgun, Being in a J-Pop Group and...
127K views • 2 weeks ago

Glen Powell on Being Trolled by His Parents and Co-Writing Hit Man
60K views • 2 weeks ago

Glen Powell Talks Twisters and Almost Getting Arrested After a Texas Longhorns...
72K views • 2 weeks ago

Biden Holds "Tense" Zoom Call with House Democrats
223K views • 2 weeks ago

CORRECTIONS Episode 108: Week of Monday, July 15
306K views • 2 weeks ago

RNC Speakers Lie About Trump's Record; Tucker Carlson Pushed for Vance to Be VP:...
1.4M views • 2 weeks ago

James Bay: Up All Night
17K views • 2 weeks ago

Daisy Edgar-Jones Went Storm Chasing with Glen Powell and Brandon Perea While...
87K views • 2 weeks ago

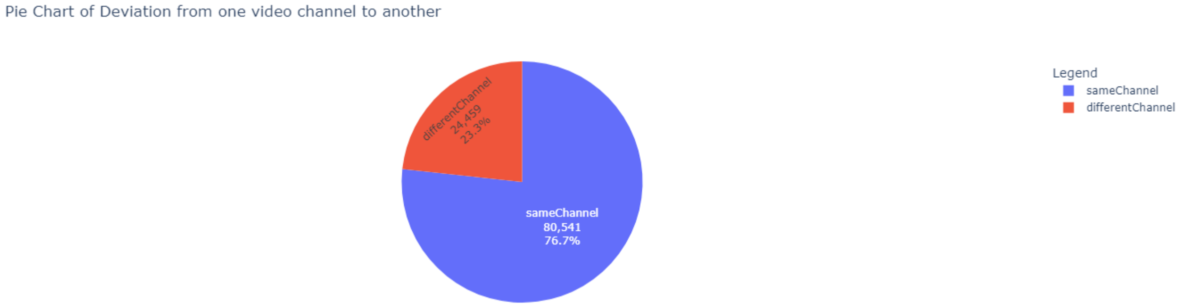
Michael Cera Talks Sausage Party: Foodtopia and Doing Voice Recordings in...
142K views • 2 weeks ago

Seth Celebrates the 8G Band's Emmy Nomination
12K views • 2 weeks ago

Trump Wants Taiwan to Pay U.S. for Protection
331K views • 2 weeks ago

Using a similar method with the category deviation, the program made a channel deviation pie Chart.

*Figure 8 - Pie Chart of Deviation from one video channel to another*



76.7% of all recommendations were from the same channel as the current video. Once again displaying an overwhelming bias to feed the user similar content.

## Corroborating "Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users"

This report analyzes the bias in YouTube's recommendation system, specifically searching for political bias. They describe that without user data it is difficult to fully analyze a recommendation system. To solve this problem, they fielded a survey of YouTube users from the fall of 2020 and recorded the recommendations they were shown. They found a small bias towards echo chambers when using user data, which grows as the user delves deeper into the recommendation algorithm. They also noted that there is stronger evidence of a platform-wide bias for more moderately conservative content. (Megan A. Brown, 2022)

Our simulation similarly found that even without user data, YouTube does not want to recommend different categories or channels from the current video with around 75 - 80% of recommendations being from the same channel / category. If user data is used to further push users towards those same categories and channels, then these numbers could grow even larger.

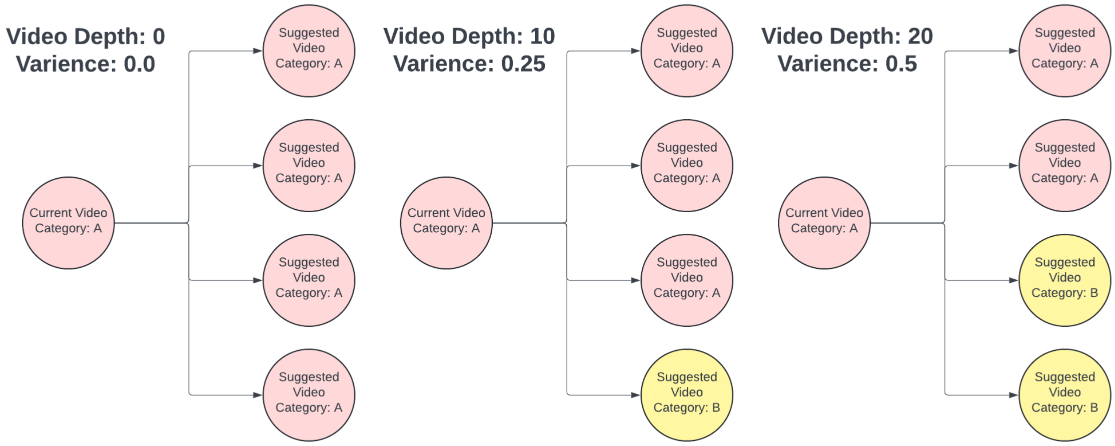
## Recommendations

Now that bias has been indicated, we can recommend methods to alleviate it. Two methods will be examined to provide options if one is to be implemented. It is also important to note that because the YouTube algorithm is a black box, it is impossible to know if these methods, or something similar, have already been implemented or not without having access. The two recommendations are: Method 1: Introducing Recommendation Variance, and Method 2: Applying a Random Surfer Approach.

### Method 1: Introducing Recommendation Variance

The plan for this approach is to increase the chance a video is recommended outside of the current category or channel the longer the user delves into that same category or channel. This is done by introducing a "variance" variable to the program. This variable is a weighting which will affect the number of videos outside the current category to be recommended. As a user continues watching a given category of video this weight continues to increase.

*Figure 9 - Diagram showcasing the recommendation variance approach.*

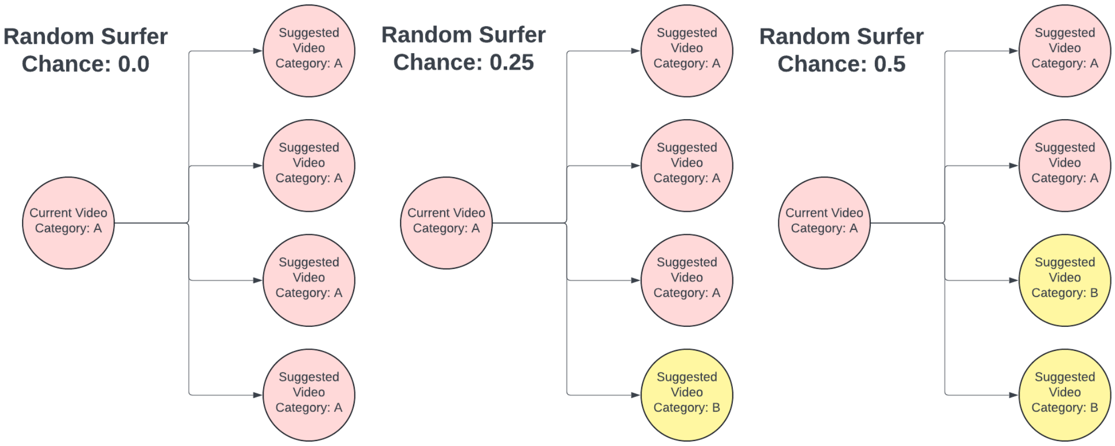

Video Depth: 0
Varience: 0.0
Current Video Category: A
Suggested Video Category: A
Suggested Video Category: A
Suggested Video Category: A
Suggested Video Category: A
Video Depth: 10
Varience: 0.25
Current Video Category: A
Suggested Video Category: A
Suggested Video Category: A
Suggested Video Category: A
Suggested Video Category: B
Video Depth: 20
Varience: 0.5
Current Video Category: A
Suggested Video Category: A
Suggested Video Category: A
Suggested Video Category: B
Suggested Video Category: B

This method will increase the chance for a user to find content outside of their current niche and lower the overall bias of the algorithm. To prevent the suggested videos from becoming completely different from the current video, it is advisable to place a maximum value for the variance variable.

### Method 2: Applying a Random Surfer Approach

The plan for this approach is to use the random surfer approach. This is a concept used graphing, specifically in page rank a search engine optimization technique, where there is a random chance that a user will arrive at a given webpage. This method can be tweaked to work with YouTube's recommendations by implementing a random chance a video is recommended outside of the current category or channel. This is done by setting a probability that every recommendation is outside of the current category or channel.

*Figure 10 - Diagram showcasing the random surfer approach.*


Random Surfer Chance: 0.0
Current Video Category: A
Suggested Video Category: A
Suggested Video Category: A
Suggested Video Category: A
Suggested Video Category: A
Random Surfer Chance: 0.25
Current Video Category: A
Suggested Video Category: A
Suggested Video Category: A
Suggested Video Category: A
Suggested Video Category: B
Random Surfer Chance: 0.5
Current Video Category: A
Suggested Video Category: A
Suggested Video Category: A
Suggested Video Category: B
Suggested Video Category: B

## Conclusion

It has been shown that YouTube has an issue with bias in its algorithm:

- 81.2% of recommendations were from the same category.
- 76.7% of recommendations were from the same channel.
- 25% of all recommendations were from the entertainment category.
- 22% of all recommendations were from the same 11 channels.

To decrease these numbers, and alleviate the bias in YouTube's system, it is recommended that at least one of these two approaches are implemented:

- **Introducing Recommendation Variance.** Introducing variance to the recommendations will decrease bias as a user continues to delve deeper into a specific category of content.
- **Applying a Random Surfer Approach.** Introducing a random surfer variable will add a constant variance to all recommendations on YouTube. This promotes varied content in all aspects of the website.

## References

Ceci, L. (2024, April 11). *Hours of video uploaded to YouTube every minute as of February 2022*. Retrieved from Statista: https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/

Elad, B. (2023, October 10). *List Of Vital YouTube Statistics Marketers Should Not Ignore In 2023*. Retrieved from EnterpriseAppsToday: https://www.enterpriseappstoday.com/stats/youtube-statistics.html

Megan A. Brown, J. B. (2022, November 12). *Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users*. Retrieved from SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4114905

# Appendix

Link to the data collection program along with all the data. Data for this report is found in the directory: “api_data/maxDepth_1000”

https://github.com/mr-rjh3/youtube-viewer