"Covid Coverage"

Exploring Youtube's Recommendation Algorithm via Breadth First Search

Zack Johnson

A Final Project for QSS 41: Analysis of Social Networks

June 2020

Introduction

I've always found Youtube addicting. Over the past few months especially, I'll click on the app as I'm getting into bed. I tell myself that I'm only going to watch one or two videos, after which I will go immediately to sleep. But on many of these nights, I find myself a few hours later, in the same position, still watching. This isn't a massive issue for me - the consequences amount to fatigue the next day and an overdose of movie trailers and basketball highlights. But in other situations you can find far worse outcomes. In 2019, *The New York Times* profiled Caleb Cain, a liberal college dropout who got "sucked into a vortex" of far-right politics on Youtube. Where I just lost sleep, Mr. Cain got radicalized into the alt-right. The article cites as a mechanism Youtube's recommendation algorithm which it says accounts "for more than 70 percent of all time spent on the site."[1]

The recommendation algorithm suggests 'related videos' based on your previous site activity. These videos show up in two places - on the homepage of the site and in the section labeled 'Up Next'. It's this 'Up Next' section that creates the aforementioned vortex. Youtube bets that if you find the video you are currently watching interesting, you'll find videos related to it interesting. There's also a nearly unlimited supply of videos available. So, with one search query you can just keep clicking on videos from the 'Up Next' section and watch interesting content for hours upon hours. Zhou et al confirms this phenomenon, finding that "the related video recommendation is the main source of views for the majority of videos on Youtube."[2] The combination of this feature's popularity and the anecdotal evidence of its influence on people makes it a ripe target for network analysis.

---

[1] Roose 2019
[2] Zhou, Renjie et al. 2010

In this paper, I present an exploratory study of the recommender algorithm as it relates to a current political issue, in this case COVID-19. I make two assumptions. First, I assume there is a broad swath of videos available on Youtube about COVID-19, with a wide variety of political orientations. Second, I assume, for a user who is interested in COVID-related content, that a typical behavior pattern is to search for a keyword then follow the chain of recommended videos to learn more. This study's aim is to understand the effect that network topology has on the type of content with which a user might interact.

In our network, we conceptualize videos as nodes and the designation of 'related' as an edge. We collect a small starting set of seed videos then 'follow' their connections through a modified breadth-first search. We then use network analysis to analyze the resulting graph. In this analysis, I will highlight 3 interesting results: 1) the second recommended video you watch seems to have the largest effect on subsequent content, 2) once you go down a content-path, it's quite difficult to backtrack, and 3) the videos you see very quickly stop being related to your original search term.

This work builds upon previous study of the recommender algorithm. The idea of 'following' the path of the recommended videos to construct a graph is well established. Airoldi et al used it to better understand music classification.[3] Schmitt et al used it to demonstrate that extremist messaging and counter-messaging often appear together on the site.[4] Korsunka also investigates the process of radicalization; starting from one video, she visualizes the jumps between genres (from mainstream to conspiracy/satire).[5] It seems also important to note that the study of recommender systems is not limited to just Youtube. There's a wealth of research

[3] Airoldi, Massimo et al. 2016
[4] Schmitt, Josephine B et al. 2018
[5] Korsunka 2020

dedicated to both improving recommender systems[6] and the effects they have on revenue on more monetized platforms.[7] It seems a widespread consensus that understanding the effect that these systems have on our daily lives is worthy of inquiry.

Data and Methods

Data was collected in Python via the Youtube Data API. All code and data for the project can be accessed here: https://github.com/zjohnson5455/covid-coverage. An overview of the API can be found here: https://developers.google.com/youtube/v3/getting-started. I specifically used the search.list method, which has parameters for both searching for a keyword and retrieving videos related to a specified video Id. I also specified that all results be in English and from the United States. This a) allows me to better understand the results (I only speak English) and b) better mimics the behavior of the type of target user I'm investigating. In this case I assume that, given how different the covid-19 response has been from country to country, a user would be primarily interested in content about their own country. I also turn safe search off so to get a fuller picture of the type of content Youtube might show you.

I start by making a generic Youtube search for 'coronavirus'. I choose this term somewhat arbitrarily - from my experience, the disease is referred to as 'covid-19' in more formal settings while 'coronavirus' is used more colloquially. Subsequent research could be done in comparing the results generated by different search terms. I limit the max results received here to 3. This a) mimics the behavior of a user who most often clicks on one of the top few results and b) ensures that we are able to follow these nodes several levels deep without generating an unmanageable amount of data. These 3 videos are added to a queue.

---

[6] Nie, Bin et al. 2014
[7] Oestreicher-Singer, Sundararajan 2012

From here, we perform a fairly standard breadth first search. At each node we visit from the queue, we use its id to find videos relating to it, again limiting results to 3. These edges are recorded and the new videos are added to the queue. We make one modification to the standard procedure. In traditional breadth first search, you would use a set to make sure you don't explore edges to nodes that have already been visited. In this case, we do still maintain a visited set and refrain from adding visited nodes to the queue again. However, we track all edges, regardless of if this new node has been explored or not. This produces an important result. Where traditional breadth first search would produce a tree structure, this modification makes it so there can be direct edges between a node and a different node higher up (or at the same level) in that tree. For the purposes of this paper, we'll refer to these edges as 'reach-back' edges. It's these reach-back edges that make the network topology worthy of study; otherwise, we would end up with 3 separate very standard trees with a branching factor of 3.

During his period of radicalization, Caleb Cain watched roughly 12,000 videos and entered roughly 2,500 search queries.[8] This would mean for each search query, he watched about 4.8 videos. So in order to mimic user behavior, the goal was to run this search 5 levels deep. Unfortunately, the Youtube Data API applies fairly strict rate-limiting. Each day, you are allotted a quota of 10,000 'points', which prevented me from finishing an exploration of the 5th level. Because search results change with time, I didn't want to explore over multiple days and introduce that variation into my dataset.

In total, I collected a dataset of 191 videos, with 284 edges between them. For each video, I record its id, title, depth at which it was found (starting with 0 for the seed nodes), and a

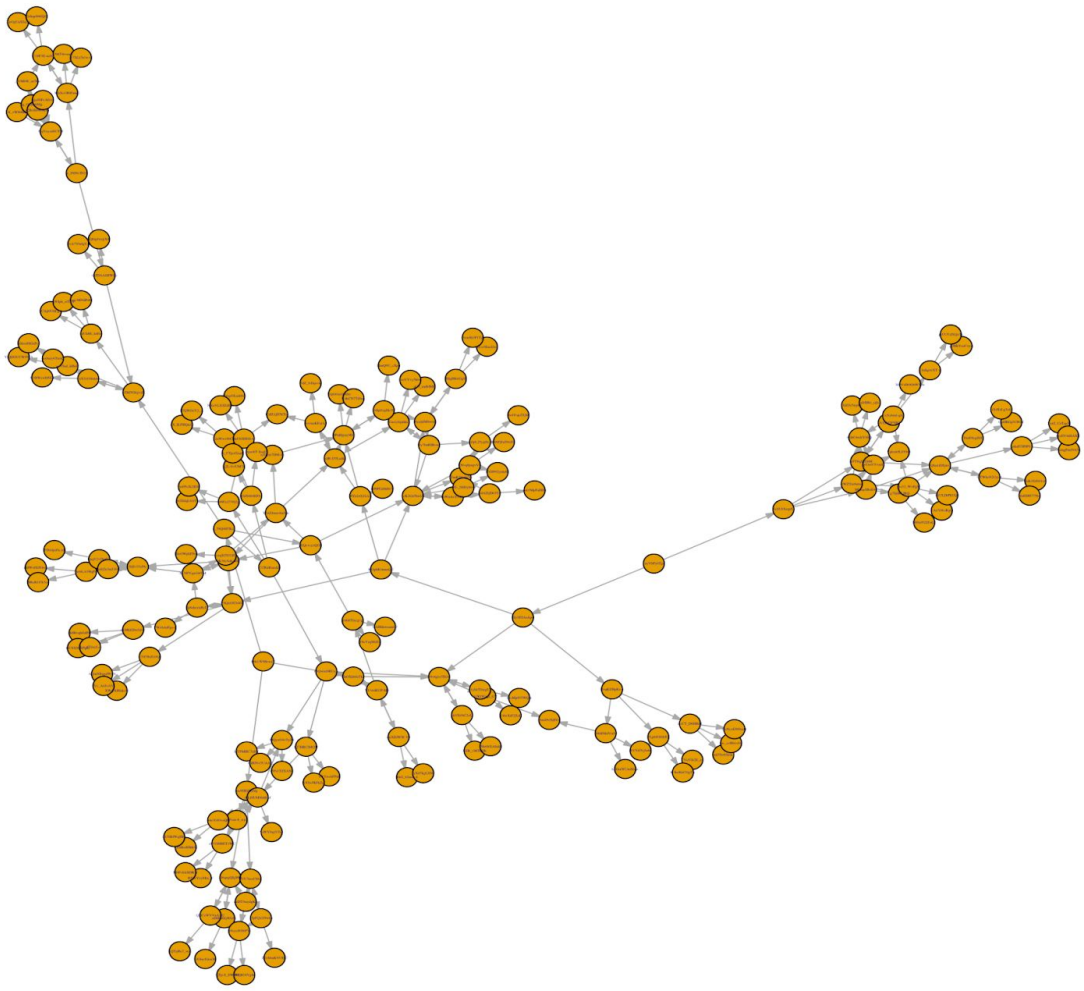
[8] Roose 2019

binary variable referring to whether the video is related to covid-19. Given the exploratory nature of this study, the procedure for determining the values of this 'CovidRelated' variable is relatively ad-hoc. I went down the list of titles and labeled each a 1 if it seemed obviously related to the virus or the economic collapse/recovery from the virus. I labeled all others a 0. Note that I did not actually watch the videos themselves, so there is of course the potential for false negatives. But for the purposes of this investigation, I assume that this false negative rate would be relatively small and would not affect the general conclusions.

All data analysis was done in R using the iGraph package. For result 1, I collect statistics on the number of components, betweenness centrality, and degree distribution. For result 2, I color the graph by depth and create mixing matrices for ties between nodes of different depths. For result 3, I color the graph by the covid-related variable and measure relative frequencies both overall and at various depths.

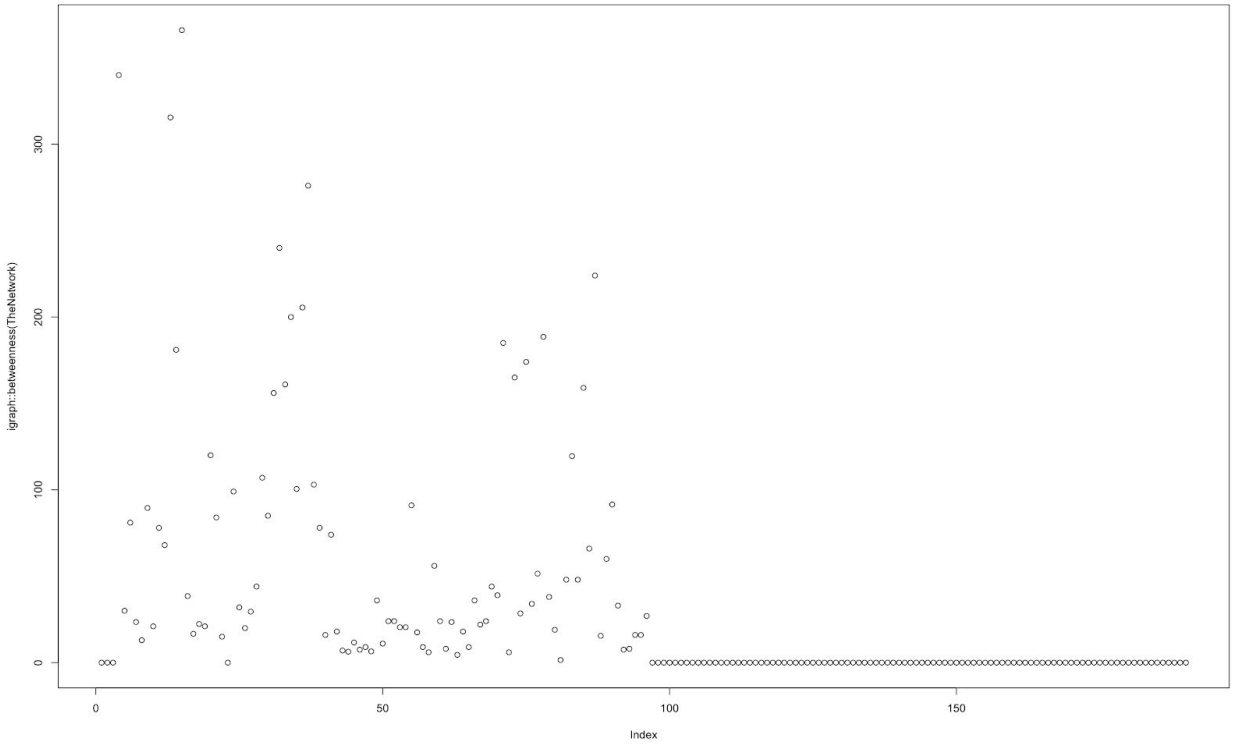
Results

1) *The second recommended video you watch has the largest effect on subsequent content*



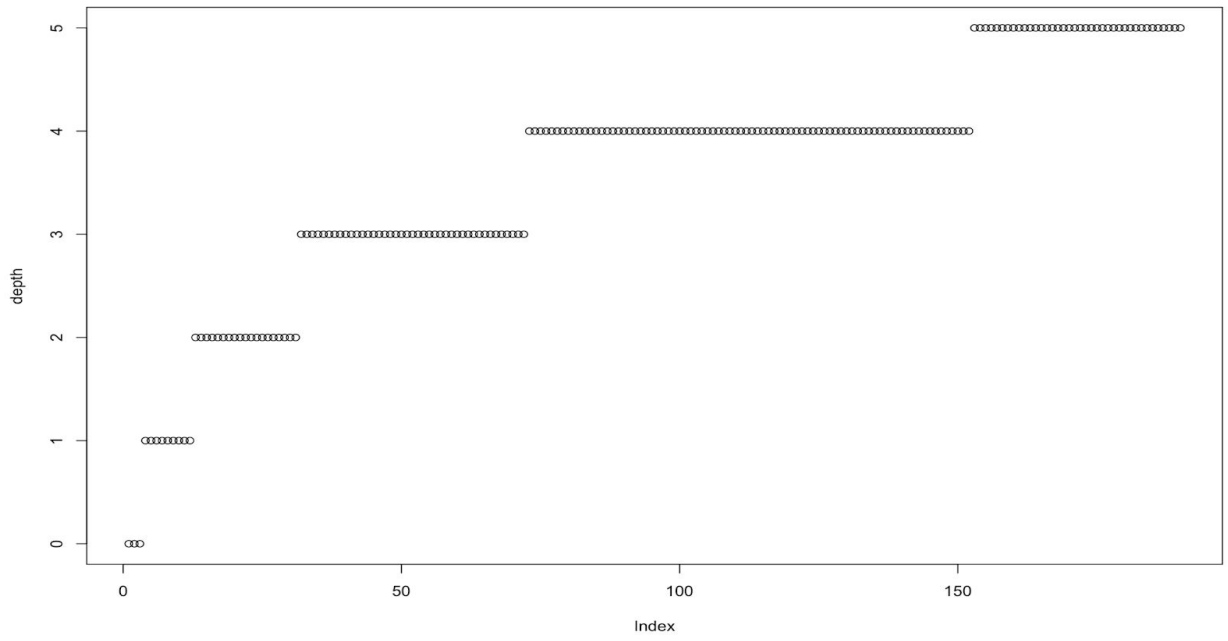
Shown above is the network for this dataset. Note that it's one giant component. This is not obvious: remember, we started with 3 separate seed nodes with no connections between them other than that they were all top results for 'coronavirus'. I see two possible explanations. First, this could be attributed to low clustering. If, starting from distinct points, we end up exploring largely the same collection of videos, that implies that there are not separate, tightly-clustered 'echo-chambers' to accidentally enter and get trapped within. Instead, it is one randomly connected hodgepodge of videos. However, looking at the graph, this isn't true. There seems to be distinct tree-like structures connected only by a few nodes in the center. This goes to the

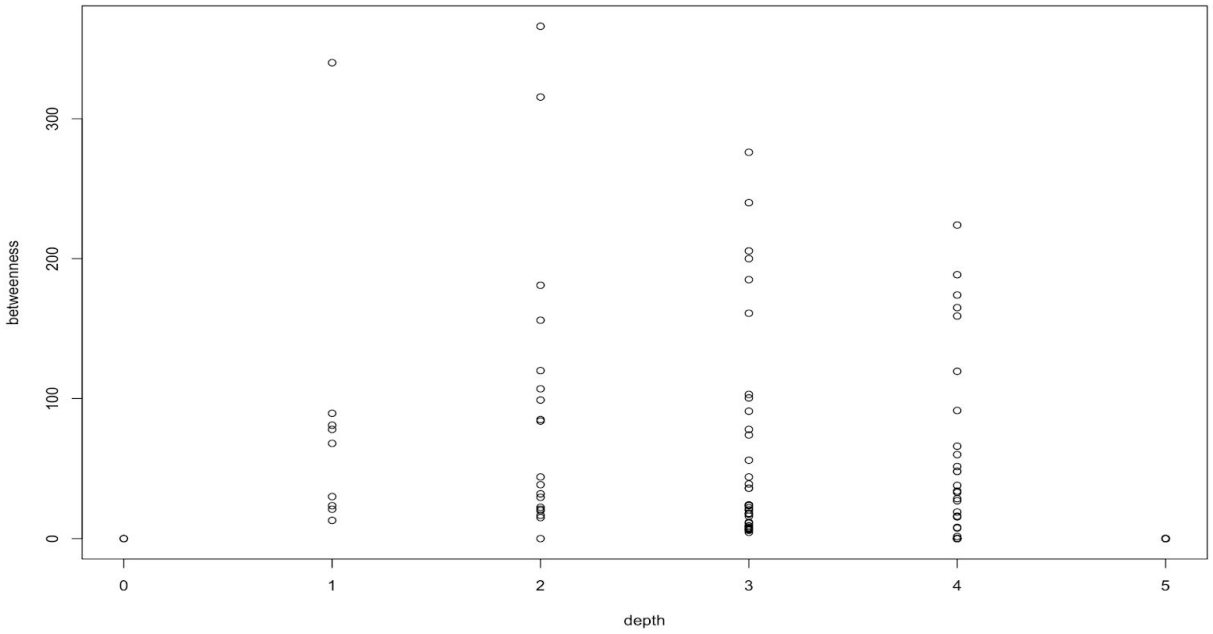
second possible explanation: that there are a few critical ‘connector nodes.’ To investigate this more formally, we can look to the distribution of betweenness centrality (shown below).



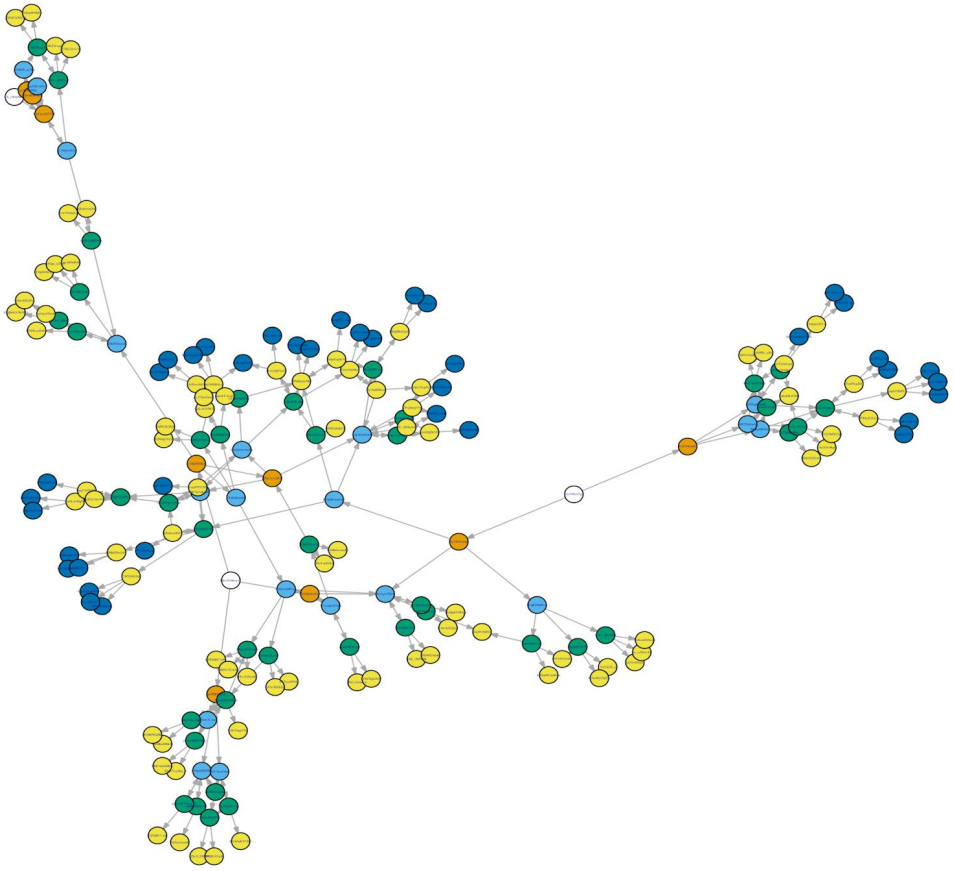
A couple of observations here. First, we have a line of nodes at the end with a betweenness score of 0. This makes sense because these nodes were added to the graph last, so have had very little time to collect edges from subsequent nodes. Second, our 3 seed nodes (the first 3 indices) also have a betweenness centrality of 0. This is interesting as it implies that our initial top search results do very little to connect different parts of the graph. Finally, the pattern of centrality appears to be almost cyclical, with nodes of high centrality occurring near each other. We also see the peaks of the “cycles” decreasing in height over time. One explanation for this might be that 1) these peaks correspond with the beginning of a new level of search depth and 2) that centrality declines with each level out. Consider the following two figures as evidence. We can compare the indices on the first figure and from the above centrality figure to

see that we are correct in #1: that the beginning of a new level out generally corresponds with a jump in centrality. However, for #2, the second figure shows us we are only partly correct. Peak centrality seems to actually occur at depth 2 and then declines from there. If we accept a high betweenness centrality as evidence of connecting two distinct parts of the graph, then nodes with the highest betweenness centrality score have the greatest affect on what type of content you see. In this case, these nodes lie on the second level of our search.





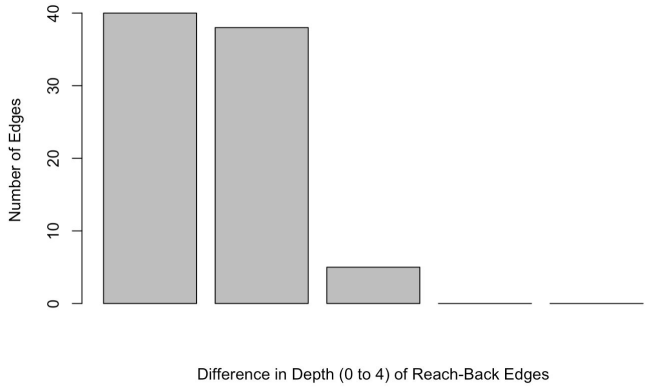
2) Once you go down a content-path, it's difficult to backtrack



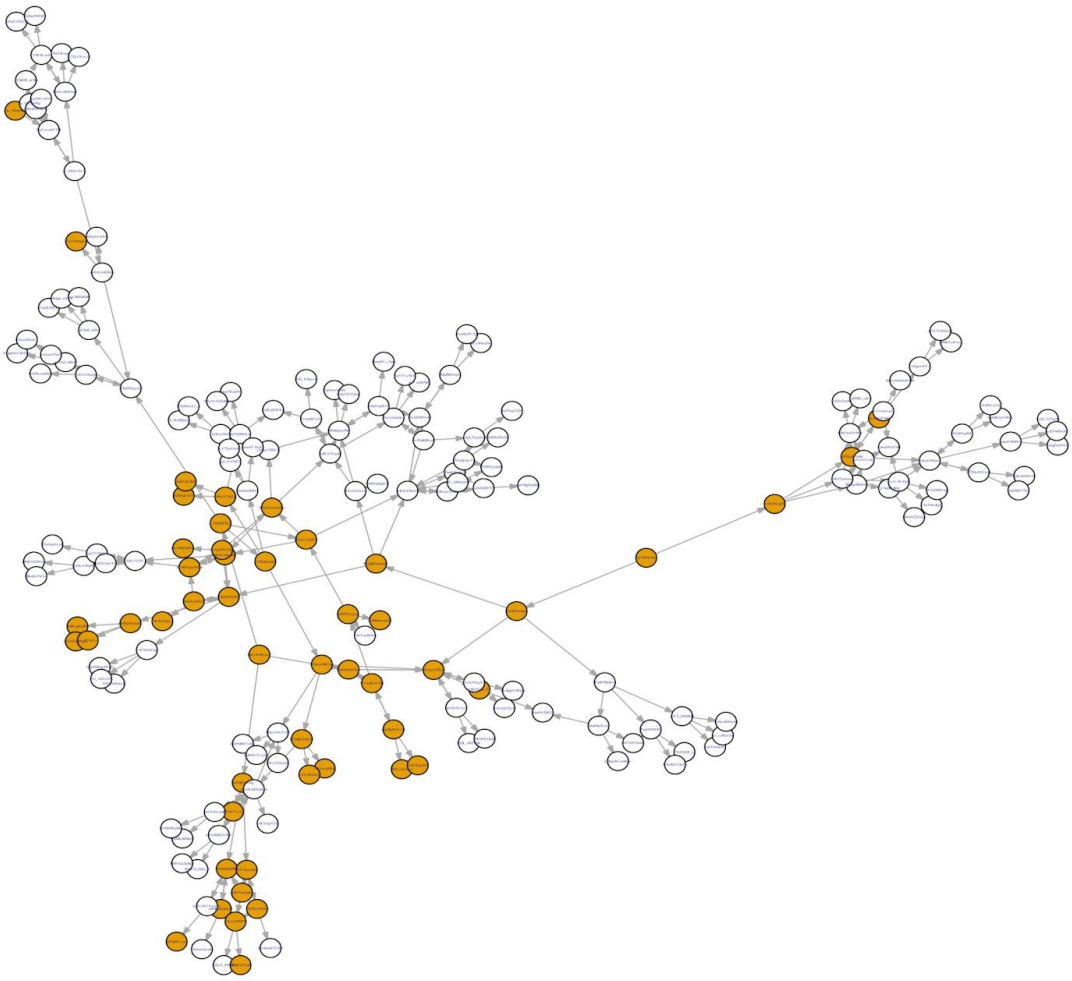
The figure above shows a coloring of the graph by depth level. Depths 0-5 correspond to the color sequence white, orange, light blue, green, yellow, dark blue. The figure below at left shows a mixing matrix of ties between levels. As we would expect, most ties are sent from level x to level x + 1. This mirrors what you would find in a traditional tree structure. All of the other edges are what I have referred to as “reach-back” edges. Note that the proportion of normal to reach-back edges is skewed for level 4, since we did not finish exploring level 5. Also remember that ties between nodes of the same level also classify as reach-back edges. Two interesting things to note here. First, there are reachback edges at every level starting at level 1. This is not a trivial result - it implies that after a threshold of about 8 nodes, we start seeing videos that you

would have encountered if you had taken a different path as well, at either the same level or an earlier one. From this we can say that it is not impossible to climb out of a content path. Second, though, nodes send reach-back edges most frequently to nodes either on its own level or on one level above. The second figure below demonstrates this phenomenon. Let's imagine a scenario in which a path of videos takes you from relatively mainstream content further and further down towards radical content. This result would imply that once you reach the radical content, it's relatively difficult to jump back to more mainstream content just by following recommended videos - you would have to perform a series of smaller jumps through multiple videos.

	0	1	2	3	4	5
0	0	8	0	0	0	0
1	0	7	20	0	0	0
2	0	7	5	45	0	0
3	0	1	15	18	86	0
4	0	0	4	16	10	41
5	0	0	0	0	0	0



3) The videos seen become quickly unrelated to the original search term



The above figure shows the graph colored by our CovidRelated variable: orange means related and white means unrelated. In total, there are 50 related nodes and 140 unrelated nodes. The figures below show a mixing matrix of ties between the categories and a breakdown of their distributions by level.

	0	1
0	173	6
1	35	69

	0	1
0	0	3
1	3	6
2	8	11
3	29	12
4	67	13
5	33	5

Note a few things here. We see the majority of ties happen between nodes of the same type: related nodes send more ties to related nodes and unrelated nodes send more ties to unrelated nodes. The proportions are also different: unrelated nodes send a higher proportion of ties among themselves than related nodes do. This implies that once a path becomes unrelated to the original subject, it will likely stay unrelated. It also implies that gradually the graph will become more skewed towards unrelated nodes as time progresses. This is furthered by the breakdown by level: the deeper down you go, the higher the proportion of unrelated nodes to related.

Conclusion

It remains true that this is an exploratory study. My dataset is relatively small and I didn't conduct anything in the way of statistical significance tests. However, I think my results here show once again that the Youtube Recommendation system is worthy of study. We have an algorithm in front of us that is incredibly popular and helps people spend hours on the site. Our result #3 tells us that this algorithm sends people down rabbit holes of content for which they did not search. Our result #2 tells us that these rabbit holes are often difficult to escape from. Luckily, our result #1 tells us that if we can influence the first few videos that a user watches,

that we can shape the type of content they interact with later on. Areas for further research might include performing a larger-scale version of this experiment, performing more explicit clustering analysis of the network, or conducting a more refined way to analyze the content of the videos encountered, perhaps with natural language processing. I may continue to be a Youtube addict for years to come. But hopefully as a society, we can minimize the production of more Caleb Cain-s.

Bibliography

Airoldi, Massimo, et al. “Follow the Algorithm: An Exploratory Investigation of Music on YouTube.” *Poetics*, vol. 57, 2016, pp. 1–13., doi:10.1016/j.poetic.2016.05.001.

Korsunska, Anna. “Network Analysis on Youtube: Visualizing Trends in Discourse and Recommendation Algorithms.” *Loretta C. Duckworth Scholars Studio*, Temple University, 22 Apr. 2020, sites.temple.edu/tudsc/2019/03/26/network-analysis-on-youtube/.

Nie, Bin, et al. “Social Interaction Based Video Recommendation: Recommending YouTube Videos to Facebook Users.” *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Apr. 2014, doi:10.1109/infcomw.2014.6849175.

Oestreicher-Singer, Gal, and Arun Sundararajan. “Recommendation Networks and the Long Tail of Electronic Commerce.” *MIS Quarterly*, vol. 36, no. 1, 2012, pp. 65–83. *JSTOR*, www.jstor.org/stable/41410406. Accessed 7 June 2020.

Roose, Kevin. “The Making of a Youtube Radical.” *New York Times*, 8 June 2019, https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html.

Schmitt, Josephine B, et al. “Counter-Messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube.” *Journal of Communication*, vol. 68, no. 4, 2018, pp. 780–808., doi:10.1093/joc/jqy029.

Zhou, Renjie, et al. “The Impact of YouTube Recommendation System on Video Views.” *Proceedings of the 10th Annual Conference on Internet Measurement - IMC 10*, Nov. 2010. *ACM Digital Library*, doi:10.1145/1879141.1879193.