Data Analytics 350 YouTube Lab

Sally Hyde and Matthew Rinker

***Abstract*— In this lab we examined data scraped from YouTube. We looked specifically at the channel VisualPolitik. To do our analysis we created networks of videos through adjacency matricies. We then examined these both innately and through the network analysis tool Gephi.**

I. INTRODUCTION

YouTube is the premier video sharing site in the world. Millions of people watch videos on this platform every day. This allows for interesting interactions between people of diverse cultures and backgrounds. In this lab, we explore the channel VisualPolitik, specifically at their English language channel. As the name suggests, VisualPolitik is a poltically focused channel. They focus on global politics and have videos on many areas of the world. In this lab we will take an in depth look at the various interactions between videos. We focus specifically on the interactions of users between videos, the level of interaction between videos of different countries, and the level of interaction between videos posted at different dates. We then used the network analysis tool Gephi to provide visual analysis of the networks we created.

II. DATA

The data we used consisted of three datasets. The first, VisualPolitik_Comments.csv was a dataset of data relating to all the comments posted on the videos we were looking at. The second, VisualPolitik_video_info.csv was a dataset of the metadata for every video. Finally, the dataset latitude_longitude.csv contained positional data for the regions in question.

We did not have many challenges related to the data. However, in the comments file there were some errors which needed to be handled. Specifically, there were some rows which had no videoId value or had an error for a videoId. This rendered them useless to us and had to be disregarded as if we could not tie a comment to a video then we had no way to aggregate the information they represented. Overall, we looked at around 250,000 comments on 286 videos. After removing the rows containing errors we were left with around 210,000 comments.

III. RESULTS

Note: All code was run on a computer with a 2 core intel i3 processor which runs at 2.2ghz. Timing values do not include the time to read in csv files and to export csv files.

To start, we created an adjacency matrix for the videos. To do this, we first created a dictionary that took each userId (user profile url) as a key. Contained in each entry was all the videoIds of the videos that user commented on. We then took that data and made a matrix with weights being the sum of users that commented on both videos. The total elapsed time for this operation was 2.24 seconds.

Next, we created another adjacency matrix this time using the countries or regions each video belonged to. For this, we opted to disregard videos tagged as having a "Global" region. To create this we required a dictionary that would take a videoId as a key and return the country or countries that the video corresponds to. We again used a dictionary for this process. We then took the information contained in our first adjacency matrix and using the videoIds and our translation dictionary created a new adjacency matrix. An important choice we made for this section was to break down compound countries into two countries for the purpose of this matrix. That means that a video with is tagged as being related to the USA and Canada would contribute to both USA and Canada separately as opposed to having a USA and Canada entry in the matrix. The total elapsed time for this operation was 0.606 seconds.

Lastly, we created an adjacency matrix for the publication data of each video. We restricted the dates to year and month pairs so that there would be greater overlap between videos for this matrix. To create this matrix we used the python date-time package to transform the publication_date column of our video info dataset to a clearer form containing only year and month. We then created another translation dictionary to take a videoId as a key and return the publication date. Finally, in the same method as when creating our country adjacency matrix we took the dictionary and our original adjacency matrix to create this new matrix. The total elapsed time for this operation was 4.66 seconds.

We then proceeded to complete some rudimentary analysis of our adjacency matrices. The results of our exploration is in figures 1 - 3.

We took the top ten entries in our matrix for each of the matrices we created.

The results from the original video matrix are in Fig. 1:

Adjacency Weight	Video 1	Video 2	Country 1	Country 2	Video 1 Views	Video 2 Views
228	Can Australia abandon Us because of Trump?	Why is there no crisis in Australia	Australia / USA	Australia	327,047	1,240,085
222	Why is Canada the most Admired Country on Earth	Why does Canada want more Immigrants?	Canda	Canada	879,632	745,368
215	Why is Canada not as Perfect as It Looks?	Why is Canada the most Admired Country on Earth	Canada	Canada	702,041	879,632
149	Why does Canada want more immigrants?	Why does Japan need more Immigrants?	Canada	Japan	745,368	1,375,526
139	Does Europe need to have its Own Army?	Poland, how is it getting Rich?	Europe	Netherlands	499,647	1,138,367
133	India vs China: War in the Himalayas?	Will India become a new Super Power?	India / China	India	451,902	893,400
130	Why does Japan need Immigrants?	Why is Switzerland so Rich?	Japan	Switzerland	1,375,526	1,422,500
120	Why do many Arab Countries ban Al Jazeera?	Why are the Saudies losing in Yemen?	Middle East	Saudi Arabia / Yemen	780,454	954,961
117	How Did New Zealand Become Rich?	Why is New Zealand the Most Prosperous Country on Earth?	New Zealand	New Zealand	334,168	296,211
113	Taiwan: The new Strategy to Defeat China?	Does Europe need to have its Own Army?	Taiwan	Europe	743,970	499,647

Fig. 1. Top Ten Entries in the Video Adjacency Matrix

The results from the Country matrix are in Fig. 2.

Adjacency Weight	Country 1	Country 2
17,982	USA	USA
8,068	USA	China
7,113	USA	Europe
5,280	China	China
3,290	USA	Japan
3,282	USA	Russia
3,200	Europe	Europe
3,126	China	Europe
2,971	USA	Saudi Arabia
2816	USA	Venezuela

Fig. 2. Top Ten Entries in the Country Adjacency Matrix

The results from the Country matrix are in Fig. 3.

Adjacency Weight	Month 1	Month 2
3,090	2018-07	2018-07
2,556	2018-08	2018-08
2,531	2018-08	2018-07
2,168	2018-10	2018-10
2,068	2017-07	2017-07
1,983	2017-11	2017-10
1,957	2018-07	2018-06
1,950	2018-11	2018-10
1,904	2017-10	2017-10
1,894	2018-10	2018-08

Fig. 3. Top Ten Entries in the Month Adjacency Matrix

From our results here, we can see some definite patterns. For example, we see that for almost every entry in our top ten list we have two related regions. For example, our top related video pair is comprised of a video relating to Australia and the USA and a video on just Australia. The same is true for the next few entries. In our top ten list for the videos we have two outliers. The 10th entry in the list is unexpected and does not follow our established patter. However, for the 7th entry in the list we can easily explain why two seemingly unrelated videos appear so high in our list. This is likely due to the fact that both videos have by far the largest view counts of any videos on our top ten list. Logically, if more people are viewing a video more people are commenting on the video. Thus, since people who are commenting on VisualPolitik videos likely watch more than one video from the channel it is clear that by sheer chance these two videos have a higher likelihood of sharing commentators because of the massive amount of people who viewed each video.

Further evidence of the trend seen from the video to video matrix is shown in our country matrix. However, here we see another interesting trend. Videos from two countries with contentious politics have a high relation to each other. This is likely due to the fact that the videos relating to these regions would likely be "hot button" topics and would be more likely to elicit a response from a viewer.

Finally, in our month to month matrix we see another trend. Videos from the same month or consecutive months have the highest weights. This is likely due to the watching patterns of viewers. If someone sees a video they are interested in enough to comment on they are likely to come back and watch more videos sooner rather than later. Thus, people are likely to comment on videos that occurred closer together chronologically. For our outliers, we can infer that there may have been some developments on a situation explained in a previous video or something happened that was related to the past video which caused users to watch one video and then the other video and comment on both.

Moving on to the network analysis portion of this lab, we used Gephi, an open-source network analysis and visualization tool. In Figure 4 is our Country-to-Country network visual, and it is evident that there is extensive cross-cultural content in VisualPolitik's video. There doesn't seem to be one country with more content than any other, and the node sizes, which indicate degree, are mostly uniform. We were tasked with also creating a video-to-video network in which we labeled the nodes with their video title (Figure 5). Despite our editing in Gephi, it is difficult to interpret this network because of the sheer amount of text. To compensate, we removed those video labels in our second country-to-country network and instead colored by region/continent (Figure 6). The legend for this visual's colors can be seen below the figure. From Figure 6, it is still evident that cross-cultural content views are the norm for VisualPolitik's channel. There are some notable nodes, particularly in the South America region, though all of the regions are well represented. It should be noted that "Global" videos refer to political videos that don't refer to a specific region. We then explored modularity within the video-to-video and month-to-month networks. In Figures 7 and 8, you can see each of the networks colored by their respective modularity classes. Interestingly, in the month-to-month network, the months

with the highest modularity classes fall from the final months of 2017 to the early months of 2018 (the blue cluster). In the future, it would be interesting to take a deeper dive into videos from these months in particular, and explore what makes them so modular.

Conclusion

From our cursory exploration into the datasets we have seen some interesting trends about user watch patterns. It is clear users who watch a video relating to one region are more likely to watch a video related to that same region or another region that has contentious politics with the first region. Additionally, users tend to watch videos that occur close together in time. That is to say that videos that are released close together chronologically are likely to have the same viewers watching them whereas videos that are published months apart are less likely to have the same users viewing them, unless they are covering something that invokes another trend (such as a topic related to something covered in the past). Our network analyses using Gephi did not disprove these findings, though it will be helpful in the future to further polish these network visualizations and use the Gephi software to its fullest potential to get more definitive results. In the future, one thing to consider is the population of VisualPolitik viewers that were captured in the scope of our data. Since we only collected data from the English-language channel, there may be many more Spanish-speaking viewers from the other version of his channel that are not accounted for. Having discovered that users tend to watch videos about their own regions, it is possible that we're undercounting views from Spanish-speaking countries in this analysis.

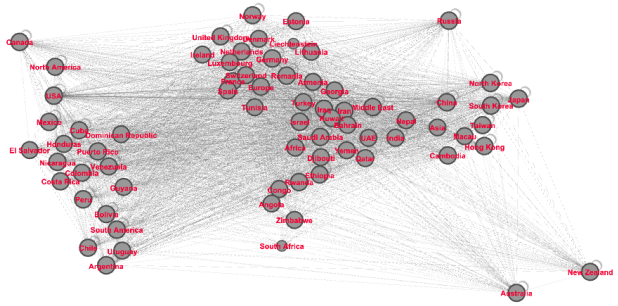

Fig. 4. Country-to-Country Network

Fig. 5. Video-to-Video Network, colored by country

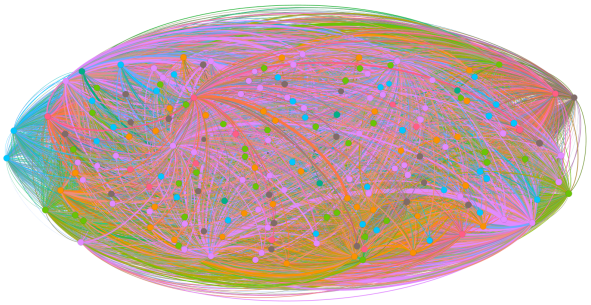

Fig. 6. Country-to-Country Network without Video Labels, colored by Region

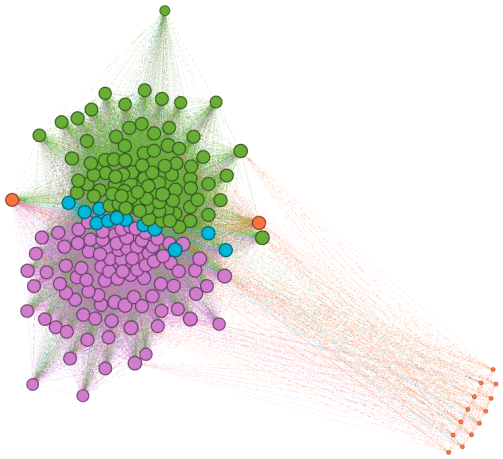

Fig. 7. Video-to-Video Network Colored by Modularity

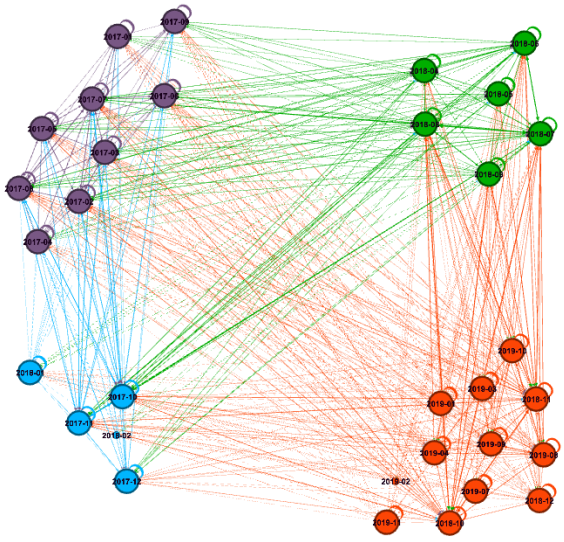

Fig. 8. Month-to-Month Network Colored by Modularity