YouTube Video Networks: "Next Video"

Alfonce Nzioka

INTRODUCTION:

In 2014 YouTube rolled out a new feature: video autoplay, where once the video currently being watched ends, instead of defaulting to a grid of suggested videos, another video starts playing automatically. The next video on the playlist is chosen from the list of recommended videos that are associated with every video. This project is an attempt to reverse engineer the algorithm that YouTube uses to decide what video to play next out of a list of related videos. I hypothesize that the choice of what video to play is the output of a ranking system that leverages the characteristics of the network of related videos.

Related videos in YouTube are an example of a field of research referred to as "recommendation systems"[1]. A recommendation system seeks to predict the "rating" or "preference" that a user would give to an item, thus recommending items to the user. There are several algorithms used in building a recommendation system. One of these approaches is called item-based collaborative filtering algorithm. This approach looks into a set of items the target user has rated and computes how similar they are to the target item *i* and then selects *k* most similar items $\{i_1, i_2, i_3 \dots i_k\}$. At the same time their corresponding similarities $\{si_1, si_2, si_3 \dots si_k\}$ are computed. Once the similar items are found, the prediction is often found using the weighted average of the user's ratings on these similar items.

This project is grounded on the concept of recommendation systems. However, the focus is not predicting how a YouTube user would rate a particular video, but rather how to rank a list of recommended videos. The idea is to generalize the calculation of similarities between each video in a set of related videos, and the current video being played. This approach is strictly video-based in that it is based on the network characteristics of a particular video and does not consider the user's playing history or ratings.

RELATED WORK:

Mislove et al. [2] studied friendship relationship between social network users. Accessible user links are crawled based on friendship between users on several online social-network websites, including Flickr, YouTube, LiveJournal and Orkut. They made a large-scale measurement on the structure of these networks. The dataset contains 11.3 million users and 328 million links. Their results confirmed the small-world and scale-free properties of online social networks. This work indicates that we should find the same properties for our network of YouTube videos. The method of obtaining data (crawling) also informs the crawling method that I use in my project.

Davidson et. al[1] describes the recommendation system used by YouTube to recommend videos to users. The set of recommended videos is generated by using a user's personal activity (watched, favorite, liked videos) as seeds and expanding the set of videos by traversing a co-visitation based graph of videos. The set of videos is then ranked using a variety of signals for relevance and diversity. The relatedness score of a video v_j to base video v_i is given as:

$$r(v_i, v_j) = \frac{c_{ij}}{f(v_i, v_j)}$$

where:

c_{ij} is the co-visitation count,

$f(v_i, v_j)$ is a normalization function that takes the "global popularity" of both the seed video and the candidate video into ac- count.

This recommendation system, however, does not work if there is no user profile. Although the methodology I propose does not generate a list of recommended videos, it ranks the relatedness of a network of related videos, regardless of the user's profile or viewing history.

METHODOLOGY:

DATA COLLECTION

My data has been collected using an YouTube crawling tool [3] which uses the "search/list#relatedToVideoId" API endpoint to retrieve videos that are related to a base video. For the purposes of this project, I used the crawling tool to retrieve videos related to 10 random videos. Each video on YouTube is uniquely identified using an 11-character unique id that comes after the "=" on the video URL. For instance, the video ID of a video whose URL is https://www.youtube.com/watch?v=WNIDcT0Zdj4 is WNIDcT0Zdj4. For each video, the crawler harvested the list of related videos, storing the unique identifier for each video, the number of comments on the video, the number of likes and dislikes, and other metadata. For each video, I obtained the video ID of the next video on the playlist. The purpose of this is to verify the accuracy of the ranking formula that I use to rank videos in the network. For each related video, the crawler harvested a list of videos that are related to that particular video.

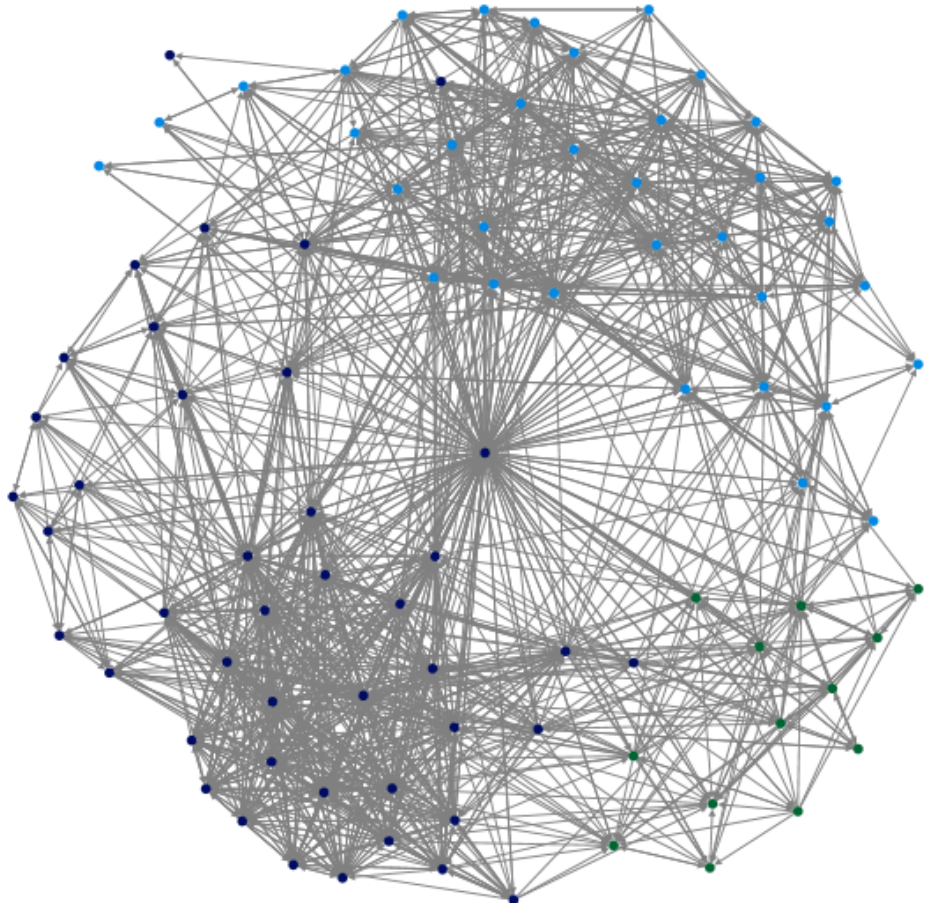
BUILDING THE NETWORK

10% of the data (videos related to a base video, randomly chosen from the 10 samples) is used to build the network while 90% of the data is used to test the ranking formula. For my purposes, I use snap to form an undirected network where each node is the video Ids from the data. An edge exists between two videos if they are related. I use NodeXL to visualize the network. Below is a preliminary network for a video whose metadata is given in the table:

Table1: Video 1 Metadata

Title	David Guetta - Hey Mama (Official Video) ft Nicki Minaj, Bebe Rexha & Afrojack
URL	https://www.youtube.com/watch?v=uO59tfQ2TbA
Number of Related videos	87
Number of Edges	853

Figure1: Video 1 network:



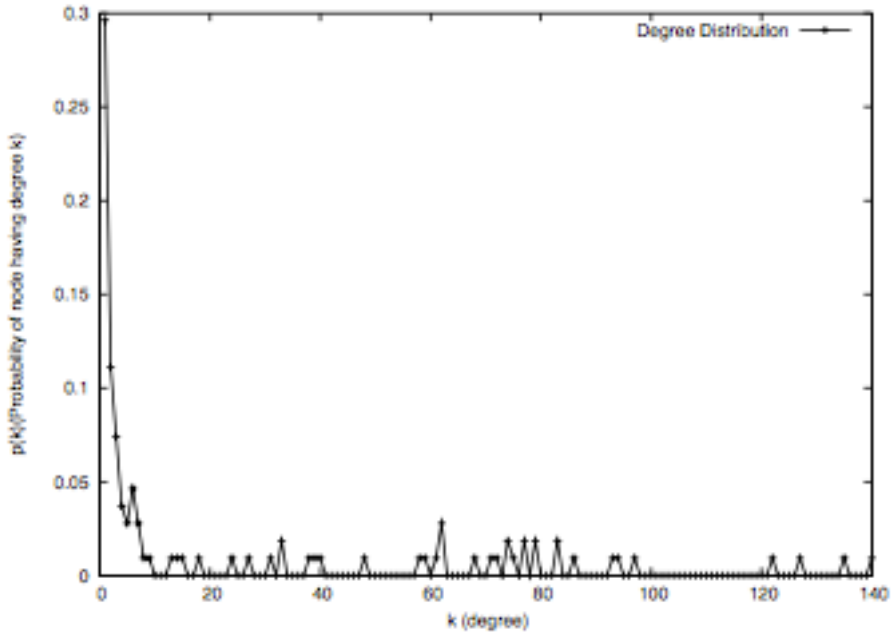
NETWORK MEASUREMENTS

Using SNAP functions, I obtain measurements to characterize the network as scale-free as well as measurements that are useful to ranking system I propose. These measurements are degree distribution, average shortest path, and clustering coefficient. To characterize the network I use

the degree distribution. The average path length, *l,* and the clustering coefficient, C, indicate whether the network has small-world characteristics.

The figure below shows the degree distribution. The distribution displays the probability *P(k)* that a given node will have degree *k.* It is easy to see that the network is not random since it follows a power-law (random networks have a Poisson distribution). The class of networks that present this type of distribution is called Complex Networks, more specifically they are defined as networks in which P (k) = k−λ. Baraba´si and Albert [4] have demonstrated that in many real networks the value of λ is invariant to the size of the network and that for most real- world networks 2≤λ≤3. For the YRN we have λ=2.3.

Figure2: Degree distribution



The average path length and the clustering coefficient tell us whether a network has small-world characteristics. The small-world phenomenon was first introduced by Milgram [5] with his famous six-degrees of separation experiment. A network is considered to be small- world if its average path length, l, grows logarithmically as a function of the number of nodes in the network. Small-world networks have a small diameter and high clustering. For the YRN we have that l ≈ 2.61 which is relatively small for the size of the network (198 nodes). The clustering coefficient, C ≈ 0.715 is much higher than an equivalent random network. The conclusion is that the YRN is a small- world network, which is important because it says that from a given video most other videos can be reached and hence considered in a recommender system.

RANKING SYSTEM

Having built the network I now use it to formulate my ranking system. The idea is to rank each video according to its importance in the network. I define a utility value for each node which expresses the importance of each video in the network.

The utility value $\mathcal{U}(n_i)$ is defined as the degree of a node n_i in relation to the entire network. That is:

$$\mathcal{U}(n_i) = \frac{\deg(n_i)}{\sum_{j=1}^{n} \deg(n_j)}$$

where *deg* is the degree of the node obtained using the GetDeg() function in SNAP. The degree of a node reflects the number of videos that are related to a particular video.

After calculating the utility value of each node, I now rank all the videos in the network according by the utility value. Preliminary results showed that the video with the highest utility value was the next video on the playlist.

References:

[1]Davidson, James, and Benjamin Liebald. *RecSys '10 Proceedings of the 2010 ACM Conference on Recommender Systems: Barcelona, Spain, September 26-30, 2010*. New York: Association for Computing Machinery, 2010. Web

[2] A.Mislove,M.Marcon,K.Gummadi,P.Druschel,andB.Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 29–42.

[3] "YouTube Data Tools." *YouTube Data Tools*. N.p., n.d. Web. 17 Nov. 2015. <https://tools0.digitalmethods.net/netvizz/youtube/>.

[4] A.-L.Baraba´siandR.Albert,"Emergenceofscalinginrandomnet- works," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999

[5] S.Milgram,"Thesmallworldproblem,"*Psychologytoday*,vol.2,no.1, pp. 60–67, 1967.