CS 6474 - Social Computing Project Midterm Milestone Report

YouTube Videos Folk Theories, Too Good to be True?

Introduction, Significance and Background

YouTube is a video sharing platform, which is used by a wide range of content creators. There is increasing interest amongst these video creators to find ways to optimize the reach of their video-based on attributes such as the usage of tags. However, the algorithms YouTube uses to curate, select and present information are opaque to users. As a result, folk theories about how these curation algorithms work have been developed. In this project, we aim to test several folk theories that try to explain why YouTube videos go viral.

The study of virality, in general, is of great importance since it allows us to glimpse at the ways in which the online and offline world affect one another. For example, it is widely believed that the 2008 presidential elections in the USA were significantly influenced by the virality of the "Yes we can" video of Obama on YouTube. The content that goes viral on any social computing platform allows us to have a better understanding of prevailing issues concerning people offline. Additionally, what goes viral on social computing systems could shape interactions within society and affect policymaking as well. We consider YouTube's trending videos as viral videos in this study (Note: we have used the words viral and trending interchangeably in the coming sections).

There are several reasons why certain posts/videos go viral on social computing systems. In this study, we aim to test folk theories related to viral videos on YouTube by evaluating the reliability of the metrics mentioned. This would help us understand for what purposes YouTube is used for (Example: If the trending videos are mainly regarding politics, then it could be logically inferred that YouTube is mainly used by its users for consuming political content). It could also help us have a better understanding of the reasons why some videos go viral, while others do not. Budding/aspiring artists are more likely to use folk theories from blogs and newspaper articles rather than read journal and conference papers on this

topic. This is why we believe that it is of importance to study the efficacy of these theories. It should be noted that it is not only individual artists who would use these, but also small businesses and NGOs. Therefore, the efficacy of these theories affects the revenue or donation income of these organizations. While, the academic community should have a discussion about reducing the barriers to access of journal articles, so that those from non-academic background can take advantage of rigorous peer-reviewed findings, we believe that it will be a long-term multi-stakeholder project and that in the mean-time, the efficacy of these folk theories than many rely on should be tested.

There have been comparable studies on virality that have been performed on other platforms such as Reddit and Twitter. A study on YouTube may provide different results from these other studies since users in many instances compartmentalize their use of social computing. By that, we mean it is possible for an individual to primarily use Facebook to connect with friends, use Twitter to consume political news and YouTube for viewing Music videos. Therefore the characteristics of virality of videos in YouTube may differ from these other social computing sites.

Previous work, in this area of research, has either looked into finding common characteristics amongst viral videos[1] and trying to use it to predict viewer counts[2] or using the survey method to see why viewers like certain types of videos[3]. While our work does involve finding the common features amongst the most viral videos, our work differs from the previous work done in two significant ways. The first is that the dataset[4] we use contains the 200 most viral videos on a daily basis for the USA (amongst 9 other countries) over a span of over 6 months and containing videos from a variety of different genres/categories. This allows our analysis to be more robust than previous studies that either used very small samples[5] or concentrated on only one genre[6]. The second and more important difference is that we are interested in testing the folk theories that are present on the internet, which to our knowledge has not been attempted yet.

Objectives and Outcomes

In our original proposal, our objectives were related to identifying the common characteristics (eg: tags and category) exhibited by the daily trending videos of YouTube, factoring in the country of origin. However, because it was suggested that our project should have more of a hypothesis-testing driven approach, we have redirected our project to testing folk theories that try to explain why YouTube videos go viral. Our original objectives were broad and tried to achieve more than what was necessary, so they have been revised to be more concise by evaluating a specific aspect related to the virality of YouTube videos.

Our original objectives involved identifying what categories/genres of YouTube videos are trending in each country irrespective of related events that are occurring or have occurred within those corresponding countries. We also wanted to determine if a correlation exists between trending YouTube videos and related events that are occurring or have occurred within those corresponding countries. Third, to identify similarities or differences in the characteristics of trending YouTube videos among several countries. And lastly, to formulate inferences regarding any other factors that could affect the virality of trending YouTube videos.

After careful consideration and discussion (in consultation with the professor and graduate teaching assistant) about the efficacy and feasibility of our proposed study, we revised our objectives. The first revised objective is to identify what are the most common folk theories related to the virality of YouTube videos from reputable sources. Two, we will be testing these folk theories by evaluating the reliability of the metrics mentioned using actual YouTube trending video statistics data.

Our original outcomes have been revised to align with our revised objectives. It was stated that data visualizations were not necessary for this project, so only simple statistical graphs and tabular data will be included instead to communicate results.

Our first revised outcome involves the use of statistical inference and analysis to state which folk theories are credible. We will point out the contradictions that exist between the various different folk theories and attempt to provide a qualitative explanation of why some folk theories work and others do not.

Description of Work Accomplished So Far

After the project proposal presentation following a meeting with Graduate Teaching Assistant Sandeep and Professor Munmun De Choudhury we followed their suggestions and planned to have a more concrete evaluation metric for our project. Hence we extensively revamped our original design and evaluation plans.

The first stage involved gathering folk theories from reliable and reputable sources. As we set about documenting the suggestions and potential challenges to testing the folk theories, we came across a few observations. One, that few of the suggestions from multiple sources, not surprisingly were repeated or had very close similarities so we grouped them together. Some advice in major blogs [9] was geared towards YouTube channels with specific size characteristics (big channels of organizations or charities). We had to generalize the advice from these sources after referring to other similar folk theories from other sources. In the end, we ended up with 12 most significant and testable theories we decided to test on.

Some folk theories were discarded based on how subjective they were or based on the possibility to codify the classification of our dataset for the feature specified in the theory. For example, Karen X Cheng in her article[8] mentions that making a video viral requires constant marketing of the video through various channels. While this seems like a perfectly reasonable assumption, testing this hypothesis will be impossible to do without surveys of the individual YouTubers of viral videos, to see how many of them have marketed their videos actively.

The shortlisted theories are present in Table 1, while Table 2 contains theories that we would like to test if time permits.

Sl. No.	Folk Theory(Hypothesis)
1	Viral Videos are short
2	The comedy genre is more likely to go viral
3	The time/day of release matters
4	There exists a positive correlation between comments and video count
5	Length of the title is short (<= 4 words)
6	Viral videos will be pre-rehearsed (rather than being candid)
7	Viral videos will have a musical quality
8	Viral videos will likely have people above the age of 25 years (or be created by those above the age of 35 years)

Table 1: Shortlisted Folk Theories(Hypotheses)

Sl. No.	Folk Theory(Hypothesis)
1	The title is written in the third person
2	For collaboration videos, see if there is a correlation between the number of views and number of subscribers for the channel
3	Viral videos run smoothly on all platforms (especially mobile)
4	Viral videos are engaging, involved and entertaining

Table 2: Additional Folk Theories(Hypotheses)

Data description

Our data set is the 'Trending Youtube Video Statistics' [4] from Kaggle. It contains the top two hundred listed daily trending videos of ten countries. Each region's data is in a separate file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count. The data also includes a category_id field, which varies between regions. To retrieve the categories for a specific video, find it in the associated JSON. One such file is included for each of the five regions in the dataset.

For data cleaning, we use Open Refine [7]. This tool was useful to check the data for duplication, checking for N/A or Null values and to get an idea about the value types and distribution of various features of our dataset and to make the decision of which were optimal features we could use for our analysis. These tools were used to unify the category_id across different countries.

Folk Theories(Hypothesis) Testing

When we analyzed the shortlisted folk theories we came across the necessity to have two separate approaches for analyzing them. Due to the nature of the theories we collected there were two methods of analysis that the dataset warranted.

Certain theories such as 'the comedy genre is more likely to go viral' did not have easy metrics to evaluate them with because the theories were either highly subjective or there was no way to codify their classification. For these theories, we decided on a manual approach to classify and test their accordance with the hypothesis. For the rest of our theories, we can codify the analyses and check the correlation of the dataset to the various folk theories we're testing.

Some examples of codifiable theories are (1) Viral Videos are short (2) Time and date of the releases matter (Monday, Tuesday are the best days) (3) There exists a positive correlation between comments and view count. (4) Title length should be short (less than 4 words).

As an example for the folk theory, which claims that Monday and Tuesday are the days which are best for YouTubers to release their videos to maximize the probability of going viral, we will use the date of publishing parameter in our dataset and chart the distribution of date of publishing vs the number of videos. Using this analysis, we can see if the theory is correct or if there exists an alternative pattern in the dataset which the folk theory missed.

Some examples of theories that need to be manually classified: (1) The comedy genre is more likely to go viral (2) Viral videos have a musical quality to them. For our manual classification, we're creating an

algorithm to select 10 of the most trending videos from each month. With 6 months of data, we end up with 60 videos. We will then divide these 60 videos among the team members and individually classify and label them for the folk theory being tested.

Description of Plan of What to do Next

Currently, we have determined the top 10 most viral videos in the USA for each of the six months of data that is provided by the dataset. The monthly top 10 videos are used only for the testing hypothesis that has qualitative metrics or the dataset does not provide the relevant quantitative metric. The videos with the highest number of views and interactions(comments, likes and dislikes) comprise the top 10 videos of the month. Videos can remain to trend longer than a single day. We have ensured that the video appears only once in the top 10 videos of the month by calculating the highest parameter values for videos that appear more than once in the trending videos list for the month.

The sixty videos obtained over the six months is split among the four team members for the manual hypothesis testing. Each member is alloted fifteen videos to evaluate. Each member watches the videos allotted to them and determines whether the video conforms with the hypothesis or not. One of the shortlisted hypotheses states that viral videos are generally pre-rehearsed and are not candid in nature. After watching the allotted video, the team member has to write a short justification why they felt the video was pre-rehearsed or candid. Having the reason will help us reevaluate the results in case of doubt or misinterpretation of the results.

We are simultaneously developing the codebase to evaluate quantitative metrics that operate on the entire database. Hypothesis concerning the correlation between the comments and views, the time of day of release of the video on YouTube and the number of words in the title is determined and evaluated. All the results are systematically tabulated for future verification purposes.

The hypothesis testing step is followed by aggregating the results from the two types of hypothesis testing discussed previously. The main goal of this step is to determine if experimental results align with the proposed hypotheses.

We will be generating simple graphs to communicate the results to the audience and also increase the interpretability of the results. We also aim to share any unexpected results that we may come across during this analysis and aggregation phase of the study.

Lastly, we will compile our work and submit a technical report and project presentation as per the requirement of the course. The presentation and report will comprise all the details of the complete project.

If time permits, we may extend the scope of the project and evaluate a few more folk theories. Additionally, we plan to carry out the same analysis for other countries in the dataset and compare the results with the USA. We would also like to identify general trends exhibited by the viral videos in different geographies.

List of Team Members

Name	**Email**	**GT ID**
Arvind Akpuram Srinivasan	arvind_s@gatech.edu	903528961
Karthik Nama Anil	kanil3@gatech.edu	903471605
Miasia J Jones	mjones386@gatech.edu	903164042
Prithvi Alva Suresh	al.prithvi@gatech.edu	903541747

Distribution of Work

Name	Responsibilities
Arvind Akpuram Srinivasan	Literature Survey; Data Collection; Formulation of Hypothesis; Hypothesis Testing; Report Documentation
Karthik Nama Anil	Literature Survey; Data Cleaning; Hypothesis Testing; Summarising & Aggregating Hypothesis Testing Results; Report Documentation
Miasia J Jones	Literature Survey; Feature Engineering for Data Analysis; Hypothesis Testing; Report Documentation
Prithvi Alva Suresh	Literature Survey; Scalable System Design; Hypothesis Testing; Report Documentation

Project Timeline

Date	Original Plan	Revised Plan
24th Sep 2019	Project proposal	Project proposal
3rd Oct 2019	-	Project feedback meeting with Prof. Munmun De Choudhury
7th Oct 2019	Literature review	Brainstorm and incorporate the feedback provided by the professor & literature review
21st Oct 2019	Design Finalization, data collection & data pre-processing	Restructure the project and reorient the goals based on the feedback, design finalization, selection of folk theories on the virality of videos on YouTube & data cleaning
30th Oct 2019	Midterm milestone report	Midterm milestone report
4th Nov 2019	Data analysis & interpreting results	Formulation of hypotheses and hypothesis testing
18th Nov 2019	Evaluation of results & visualisation	Aggregation the results of hypothesis testing, validate the hypothesis & summarising the results
1st Dec 2019	Final project presentation & development	Final project presentation & development
9th Dec 2019	Final project report	Final project report

Reference

[1] Feroz Khan, G. and Vong, S. (2014), "Virality over YouTube: an empirical analysis", *Internet Research*, Vol. 24 No. 5, pp. 629-647.

[2] D. A. Shamma, J. Yew, L. Kennedy, and E. F. Churchill, "Viral actions: Predicting video view counts using synchronous sharing behaviors", *ICWSM*, 2011.

[3] Guadagno RE, Rempala DM, Murphy S, Okdie BM, "What makes a video go viral? An analysis of emotional contagion and internet memes", *Comput Hum Behav* 2013, 29(6):2312–2319

[4] Trending YouTube Video Statistics. (2019). Retrieved on 26 September 2019, https://www.kaggle.com/datasnaek/youtube-new

[5] West, T. (2011). "Going viral: Factors that lead videos to become internet phenomena." *Elon Journal of Undergraduate Research*, 2(1), 76-84.

[6] De Choudhury, M., Sundaram, H., John, A., & Duncan Seligmann, D. (2009). "What makes conversations interesting?" *Proceedings of the 18th International Conference on World Wide Web - WWW 09*

[7] OpenRefine: A free, open-source, a powerful tool for working with messy data. Retrieved from: http://openrefine.org

[8] Cheng K.X (2013, July 31) "10 ways to make your video go viral". Retrieved from: https://medium.com/this-happened-to-me/10-ways-to-make-your-video-go-viral-d19d9b9465de

[9] Mossaver M (2015, Nov 5) "Secrets of YouTube – what makes a video go viral". Retrieved from: https://www.theguardian.com/voluntary-sector-network/2015/nov/05/youtube-what-makes-video-go-viral-charities