International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 1, January - February 2017 **ISSN 2278-6856**

Video Popularity Distribution and Propagation in Social Networks

[1]Subhankar Ghosh , [2]Selva Kumar S.

[1]Computer Science and Engineering BMS College of Engineering Bengaluru, India

[2]Assistant Professor Computer Science and Engineering BMS College of Engineering Bengaluru, India

Abstract

Online Social networks (OSNs) have become a major platform for communication as well as for distribution of content such as images, videos and external links to blog posts. Content distribution in self-hosted websites like YouTube, Vimeo, etc. generally conform to the Pareto principle where 20 percent of the contents account for 80 percent of the views. Yet the popularity dynamics of videos being shared in OSNs, from Video Sharing Sites (VSSes) is mostly unknown. The interaction of these contents with social networks dramatically increases the skewness of their popularity. To this end, we discuss the popularity dynamics of videos in VSSes focussing on views, ratings and comments. We then present an emulator which replicates the viewing and sharing behaviours of users in an OSN. Our model is based on the Galton-Watson branching process, incorporating the effect of the sharer's influence in the propagation of videos through the network. Simulation results show high unevenness in the distribution and properly captures the popularity dynamics of videos in OSNs.

I. INTRODUCTION

In the past few years, social networks have dramatically accelerated the process of information propagation in the form of blog posts, images and videos. With the advancement in broadband technologies large amounts of data can be accessed and shared online. Videos have thus become an important type of media content spreading over the internet. As of 2016, there are over 1.3 billion YouTube users and more than 4.9 billion videos being viewed on the website everyday [1]. Simultaneously, over 300 hours of video content are uploaded to YouTube on a daily basis [2]. Usually users watch videos on VSSes like YouTube and Vimeo, which host the data on their own servers. The dynamics of video popularity has dramatically changed since the introduction of the sharing feature. Now videos on these websites can be shared to various social media networks such as Facebook, Twitter, Reddit, etc.

Table I Video metadata

video ID	11-digit unique string
uploader	uploader's username
age	days the video has been on YouTube
category	video category
length	video length
views	number of the views
rate	video rating
ratings	number of the ratings
comments	number of the comments
related IDs	up to 20 related video IDs

Videos in VSSes have a distinctive feature over those in OSNs. Though videos in OSNs are primarily hosted by VSSes, they propagate via friendship links upon sharing. On the contrary, in VSSes, videos are discovered from the featured, popular and related videos lists. Understanding these distinctions provide valuable information to ISPs, CDNs and website administrators.

Privacy protection in OSNs restrict us from crawling the websites and thus it is difficult to obtain a rich dataset for proper analysis. To this end we analyse the video views, ratings and comments on the videos uploaded to YouTube on 3rd March, 2007. We study the popularity evolution of the videos over a period of one month. The results show a high unevenness in video views distribution with less than 10 percent of the videos gathering over 85 percent of the views.

This skewness in popularity distribution becomes more pronounced with time. We then present an emulator which replicates the video viewing and sharing behaviour of users in OSNs. We have used the Galton-Watson stochastic branching process for building the model. Simulation results make it evident that the recommendation strategy in OSNs amplify the skewness of popularity of videos shared in the network. In Section II we discuss the related works on the topic. Section III presents our analysis of the YouTube dataset. We present the emulator in Section IV and the simulation results in Section V. Finally, we conclude with possible future works in Section VI.

II. RELATED WORKS

There have been significant studies discussing the workload on media servers focussing on access locality and the popularity of online videos [3][4][5][6]. The authors in [7] studied the user behaviour and popularity distribution of videos hosted on MSN. Though MSN hosts very few videos and lacks a social networking aspect, the analysis presented by the authors resulted in a VoD design which significantly reduced the bandwidth costs incurred by the servers. Halvey et al. [8] presented the social networking aspect of YouTube. In [9], the authors studied

various social networks and verified the small-world and power-law features of social networks.

Cheng et al. [10] crawled the YouTube website for a period of four months and presented an in-depth study on the statistics of YouTube videos ranging from video length distribution to active life span.

Most of the previous works on information propagation are based on epidemic models like the Susceptible-Infectious-Recovered (SIR) [11][12]. A similar work [13] was performed on Flickr to study the propagation of photos which have a longer active life span compared to videos. The authors in [14] studied the factors that affect information dissemination (and its structure) by examining large scale email data. They concluded that the popularity depends on the social and organizational contexts as well. Rodrigues et al. [15] presented the effects of the 'word-of-mouth' technique on information dissemination using data gathered from Twitter. They studied the height, width and size of the propagation trees. Li et al. [16] studied the propagation structures and presented the S^2I^3 model which incorporates the viewing and sharing behaviours of the users.

Major research works have been performed focusing on viral marketing. Budak et al. [17] presented a cascade model to stimulate the spread of good information while simultaneously restricting the spread of bad information.

Bakshy et al. [18] concluded that the largest cascades of propagation are generated by users who have the greatest influence, e.g. number of followers.

A model for simulating video propagation in OSNs was presented by the authors in [19] which focussed on the effect of video properties on the popularity distribution.
Our work incorporates the effect of the initiator/sharer's influence along with the video's ratings

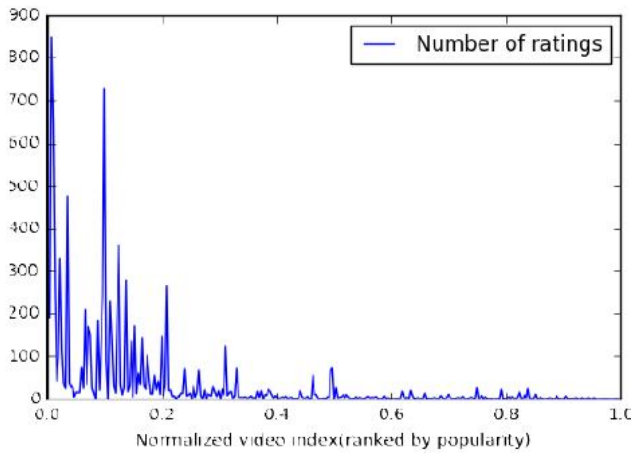

Fig. 1. Ratings

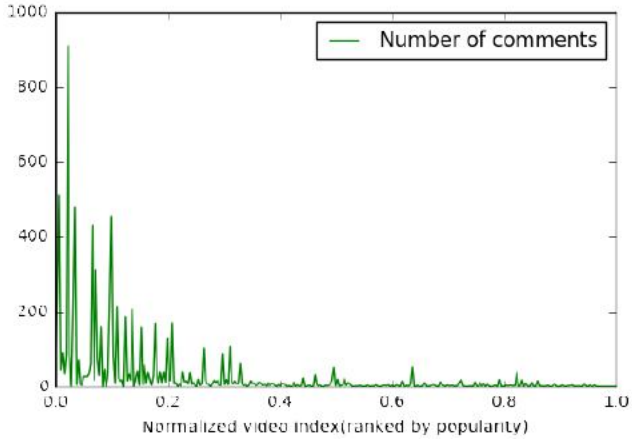

Fig. 2. Comments

on the dissemination of video content in the social networks.

III. ANALYSIS OF POPULARITY MEASURES

We analyze the statistics of YouTube videos obtained from an online dataset. The dataset provides the information presented in Table I. The data crawled provides measures such as the number of ratings, comments, views as well as the user rating for the video. This attributes are essential for analysing the videos' popularity. We take a closer look at the videos uploaded on March 3rd. The crawl results after two days (March 5th), one week (March 12th) and one month (April 2nd) give an idea of the popularity dynamics. Fig. 1 and 2 plot the number of ratings and number of comments on the videos, as observed on April 2nd, respectively. The videos are ranked based on the number of views.

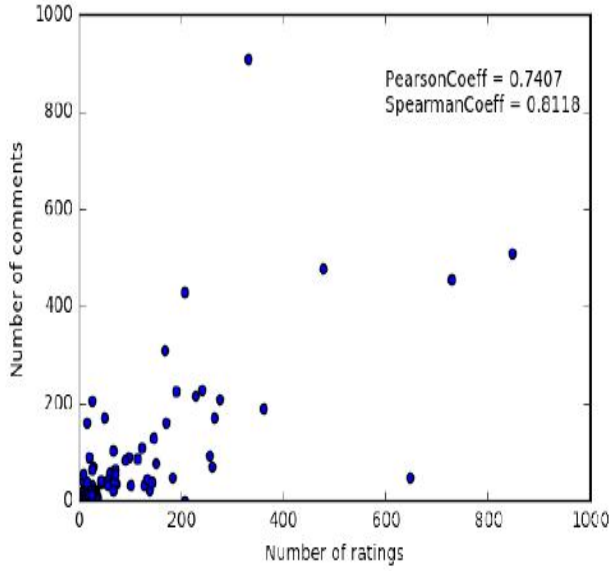

Fig. 3. Relationship between number of ratings & number of comments

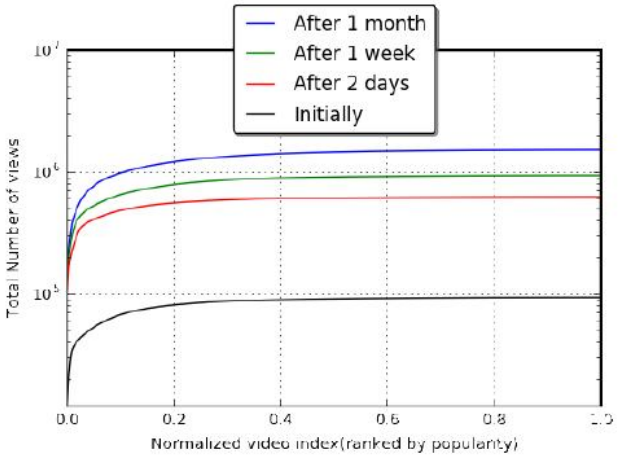

Fig. 4. Video views

Fig. 3 shows the correlation between the two quantities. We calculate the Pearson correlation coefficient (ρ_p) to be 0.7407 and the Spearman's correlation coefficient (ρ_p) as 0.8118. The quantities show strong correlation with the popular videos gathering more ratings and comments. We do find a few outliers and instances where a highly ranked video has never been rated or commented on.

The Pareto principle is generally used to describe the unevenness in distributions. The plot of video views vs ranking, in Fig. 4, presents the popularity distribution of the videos in our dataset. From the plot, we can observe that less than 10 percent of the popular videos gather majority of the views. This skewness in distribution persists even after a month and becomes even more prominent with time.

IV. MODELLING VIEWING AND SHARING BEHAVIOURS

We have used a modified form of the Galton- Watson stochastic branching process to emulate the viewing and sharing behaviours of users in Online Social Networks. The emulator can generate synthetic user requests to a collection of videos (from YouTube), rate the videos and share it among the adjoining nodes (friends). It can be used to study the popularity dynamics of videos in social networks and also for the purpose of optimizing video caching algorithms.

A. Methodology

We study the effect of the 'word-of-mouth' technique on the popularity dynamics of videos in social networks. To this end we have used the data crawled from YouTube and simulated the propagation through a synthetic social network. The network is an Erd˝os-R´enyigraph consisting of randomly generated nodes. A section of the network obtained during a sample propagation is presented in Fig. 5. Videos in the dataset have the attributes of ID, Uploader, age of the video, category, length in seconds, the number of views (*nviews*), video rating (on a scale of 5), number of ratings (*nratings*), number of comments (*ncomments)* and a list of recommended videos. The nodes of our network represent users with five attributes. The user ID, viewing rate of media content(V*wRt*), number of friends (*nf*) and number of uploads (*nu*).

$$InfSc_k = \frac{\log(1+nf_k)+\log(1+nu_k)}{\log(1+max_{\text{nf}})+\log(1+max_{\text{nu}})+b} \quad (1)$$

Using these attributes we give each user an Influence Score $(InfSc)$which determines his influence on his friends. The value of $(InfSc)$depends on the number of friends and uploads made by the user till date as in eqn. 1. Thus, a user with more friends and uploads is considered more influential. The probability of a video being viewed by a user depends on the Influence Score of the initiator/sharer and the prevaling video rating.We have normalized the score of each user by the greatest value of each of the two attributes in the dataset. Thus the range of values of $(InfSc)$lies between 0 and 1.

$$E_i = \sum_{k=1}^{S} InfSc_k \sum_{j=1}^{D^k} (VwRt_j) \quad (2)$$

An expected number of views (eviews) is assigned to each video which represents the maximum number

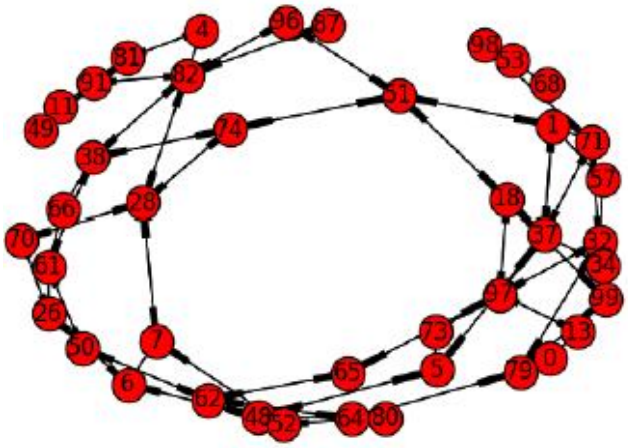

Fig. 5. Subgraph obtained during a sample propagation

of views it could accumulate irrespective of the actual viewing behaviour in the network. We assign an expectation (E_i) to each video. E_i (in eqn. 2) is the difference between the expected number of views and the actual number of views (*nviews*). S represents the total shares of a video *i*. D^k refers to the number of friends a particular sharer *k* with influence *Inf Sck* has. *VwRtj* is the viewing rate of user j with respect to the total number of videos appearing on his feed.

$$P_i = \frac{E_i - nviews_i}{\sum_{v=1}^{V} E_v - nviews_v} \quad (3)$$

A recommendation probability (Pi), which is the expectation of a video upon the total expectation for all the videos in the system, is calculated for each video. A larger value of Pi (in eqn. 3) denotes that a user request will have greater probability of being assigned to that video. nviewsi gives the actual number of views of video i, while V is the number of videos in the system. Our model assumes that only the videos shared by a friend, and thus appearing in the news feed of the user, will be viewed and shared.

B. Implementation

Algorithms 1 presents an implementation of our emulator. A new user request is assigned to a video i according to the recommendation probability (Pi) values. The initiator introduces the video into the network and further sharing behaviour by the friends causes the propagation. A video is viewed if the product of its rating, the user's viewing rate and the parent's influence score is greater than the viewing threshold value (V wRthr). Similarly, an increment in the number of shares of a video occurs when the product of its rating and the user's sharing rate and influence score.

Algorithm 1 Emulator for video viewing and sharing behaviours in OSNs

```
for request = 1 to N do
    a video request is generated by a user;
    request is assigned to video i with probability P_i;
    initiator is added to queue and to connected;
    for user in queue do
        if user is initiator then
            nviews_i++;
            eviews_i++;
            update VidRt_i incorporating user response;
            add friends to queue and to connected;
            increment eviews_i by number of neighbors;
        else
            if VidRt_i * VwRt_user * InfSc_parent ≥ VwRthr then
                nviews_i++;
                eviews_i++;
                update VidRt_i incorporating user response;
            end if
            if VidRt_i * ShRt_user * InfSc_parent ≥ ShRthr then
                add neighbors to connected if not present already;
                add neighbors to queue if not in connected;
                increment eviews_i by number of neighbors;
            end if
        end if
    end for
    update E_i and P_i
end for
```

surpass the sharing threshold (ShRthr). For the value of InfScparent, we consider the first sharer in situations where multiple users have shared the same video. We dynamically update the popularity of the videos after each complete propagation through the network.

V. RESULT

Fig. 6 presents the simulation result of the emulator's performance. The result of the propagation shows a high skewness in the number of views. We see that less than 10 percent of the videos attribute for the majority of the total number of views. This states that the 'word-of-mouth' technique results in greater unevenness in the viewing behaviour of media content in social networks as compared to the 80-20 distribution of the Pareto principle. It can also be noted that the unpopular videos on a video sharing site are not deleted and can accumulate more views slowly over time. On the contrary, videos in social networks are completely lost if they are not re-shared and thus cannot gather any further views over time.

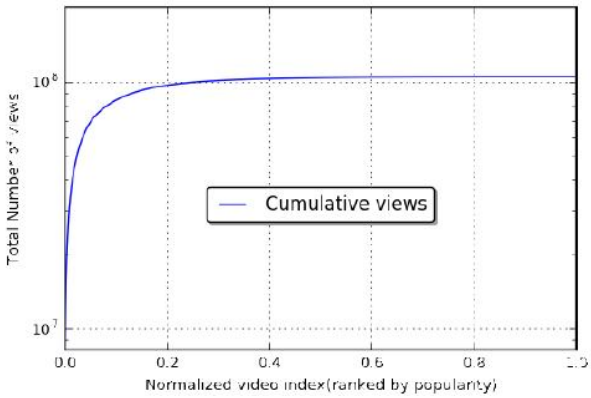

Fig. 6. Simulation results of emulator

VI. CONCLUSION

In this article we discussed the various popularity metrics of videos shared in OSNs, which are primarily hosted in VSSes like YouTube. We saw that most videos gain very few comments and ratings compared to their number of views. We discussed the skewness of the video popularity based on the observation that less than 10 percent of the videos acquire the majority of views and this skewness persists even after an extended period of time. Finally, we presented an emulator to model the video viewing and sharing behaviours of users in OSNs. Our emulator captures the unevenness of video popularity in OSNs considering features such as the influence of the sharer and the prevailing popularity of the video.

There are many possible future works relevant to our study. Incorporating the effect of the social structure, in tandem with the discussed parameters, on the video popularity would be an interesting prospect. We expect to further improve our model in our future works, for optimal usage by VSS operators.

REFERENCES

[1]. http://www.statisticbrain.com/youtube-statistics/
[2]. http://fortunelords.com/youtube-statistics/
[3]. W. Tang, Y. Fu, L. Cherkasova, and A. Vahdat, Long-term Streaming Media Server Workload Analysis and Modeling, HP Labs, Tech. Rep., 2003.
[4]. S. Acharya, B. Smith, and P. Parnes, Characterizing User Access To Videos On The World Wide Web, in Proc. Of ACM/SPIE Multimedia Computing and Networking, 2000.
[5]. J. M. Almeida, J. Krueger, D. L. Eager, and M. K. Vernon,Analysis of Educational Media Server Workloads, in Proc. Of ACM NOSSDAV, 2001.
[6]. H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, Understanding User Behavior in Large-Scale Video-on-Demand Systems, ACM SIGOPS Operating Systems Review, vol. 40, no. 4, pp. 333344, 2006.
[7]. C. Huang, J. Li, and K. W. Ross, Can Internet Video-on-Demand be Profitable? in Proc. of SIGCOMM, 2007.
[8]. M. Halvey and M. Keane, Exploring Social Dynamics in Online Media Sharing, in Proc. of the WWW Poster Paper,2007.
[9]. A. Mislove, M. Marcon, K. P. Gummadi, P. Dreschel, and B.Bhattacharjee, Measurement and Analysis of Online Social Networks, in Proc. of ACM IMC, 2007
[10]. X. Cheng, C. Dale, and J. Liu, Statistics and Social Network of YouTube Videos, in Proceedings of the 16th International Workshop on Quality of Service (IWQoS08). 229238.
[11]. A. Ganesh, L. Massoulie, and D. Towsley, The Effect of Network Topology on the Spread of Epidemics, in Proc. Of INFOCOM, 2005.
[12]. R. Pastor-Satorras, and A. Vespignani, Epidemic Spreading in Scale-free Networks, Physics Review Letters, 2001.
[13]. M. Cha, A. Mislove, and K. P. Gummadi, A Measurementdriven Analysis of Information Propagation in the Flickr Social Network, in Proc. of WWW, 2009.
[14]. D. Wang, Z. Wen, H. Tong, C.Y. Lin, C. Song, and A.L. Barabasi, Information Spreading in Context, in Proc. Of WWW, 2011.
[15]. T. Rodrigues, F. Benvenuto, M. Cha, K. P. Gummadi, and V. Almeida, On Word-of-Mouth Based Discovery of the Web, in Proc. of IMC, 2011.
[16]. H. Li, X. Cheng, and J. Liu, Understanding Video Sharing Propagation in Social Networks: Measurement and Analysis,
[17]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 10, 4 (2014), 33.
[18]. C. Budak, D. Agrawal, and A. E. Abbadi, Limiting the spread of misinformation in social networks, in Proceedings of the 20th International Conference on World Wide Web (WWW11). 665674.
[19]. E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, Everyones an influencer: Quantifying influence on twitter, in Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM11).
[20]. H. Li, J. Liu, K. Xu, and S. Wen, Understanding video propagation in online social networks, in Proceedings of the 20th IEEE International Workshop on Quality of Service (IWQoS12).
[21]. P. Erdős and A. Rényi, On Random Graphs, Publ. Math. 6, 290 (1959).
[22]. E. N. Gilbert, Random Graphs, Ann. Math. Stat., 30, 1141 (1959).