Content Popularity Evolution in Online Social Networks

Flavio Figueiredo Fabricio Benevenuto Jussara M. Almeida
{flaviov, fabricio, jussara}@dcc.ufmg.br
Universidade Federal de Minas Gerais (UFMG)

ABSTRACT

Understanding content popularity growth on Online Social Networks (OSNs) is of great importance to Internet service providers, content creators and online marketers. However, most previous studies of OSNs are based on static views of the system, thus neglecting the temporal evolution of the network, and a possible correlation with content popularity growth. Moreover, previous analyses also greatly neglect the impact of the referrers (i.e., incoming links from external sites) on content popularity. We here provide some initial results on the analysis of content popularity growth in YouTube videos. Our study is based on three video datasets, namely popular videos, randomly collected videos, and copyright protected videos, with distinct characteristics in terms of temporal popularity evolution. We also characterize the different referrers that most often lead users to YouTube videos. Our results shed some light into aspects that impact content popularity growth.

Categories and Subject Descriptors

C.4 [**Computer Systems Organization**]: Performance of Systems—*Measurement techniques*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Human Factors, Measurement

Keywords

OSNs, YouTube, video popularity, popularity growth

1. THEORETICAL BACKGROUND

This paper describes a PhD work, being developed at the Universidade Federal de Minas Gerais. The work started in July 2010 and is expected to finish by July 2015.

Given that Online Social Networks (OSNs) are currently a major segment of the Internet, understanding content popularity growth on these networks is of great relevance to a broad range of services, from technological, economical and social perspectives. Such understanding can drive the design of cost-effective caching and content distribution mechanisms as well as uncover potential bottlenecks in system components such as search engines [5]. Moreover, predicting popularity is also important not only for supporting online and viral marketing strategies as well as effective information services (e.g., content recommendation and searching services) but also because it may uncover new (online and offline) business opportunities. From a sociological point of view, a deep study of popularity evolution may also reveal properties and rules governing collective user behavior [6].

2. OBJECTIVES

The main objective of our work is to understand the diffusion and evolution of content popularity in large scale OSNs. In particular, we are interested on dealing with OSNs which focus on user created content (UGC) [1], due to the volume [5, 6] and more complicated nature of such media [4]. One representative example of such OSNs is YouTube[2], being the largest video sharing network nowadays.

In broader terms, we aim at understanding the evolution of content popularity with respect to three main research challenges (RC): (1) popularity growth patterns, which are related to the different patterns of popularity evolution across UGC content; (2) the referrers (i.e. incoming links) of UGC, which deals with how users find content on OSN and how this impacts popularity evolution; and finally, (3) how changes in the structure of the OSN affect popularity.

We begin our study with a review of the related literature in Section 3. A description of each challenge is presented in Section 4, while our research methodologies are presented in Section 5. In order to provide initial insights on *RC 1-2*, we characterized the growth patterns of video popularity on YouTube [8]. Using newly provided data by the application, we analyzed how the popularity of individual videos evolves since the video's upload time. We also characterize the different referrers for each video. Our results reveal differences in popularity evolution patterns depending on different video samples (top, random and copyrighted). These are presented in Section 6. Section 7 concludes the work.

3. RELATED WORK

Static views of popularity: There have been a few studies that address content popularity on OSNs, and, particularly, on video sharing systems. Cha *et al.* [5] presented

[1]Online radios, such as LastFM (`http://www.last.fm`), are examples OSNs which does not deal with UGC.
[2]`http://www.youtube.com`

WebMedia'11: Proceedings of the 17th Brazilian Symposium on Multimedia and the Web. XI Workshop on Ongoing Thesis and Dissertations.
October 3 -6, 2011, Florianópolis, SC, Brazil.
ISSN 2175-9650.
SBC - Brazilian Computer Society

an in-depth study of two video sharing systems. They analyzed popularity distribution, popularity evolution and content characteristics of YouTube and of a popular Korean video sharing service, investigating mechanisms to improve video distribution, such as caching and P2P distribution.

Gill *et al.* [9] characterized the YouTube traffic collected at the University of Calgary campus network, comparing its properties with those previously reported for Web and streaming media workloads. In particular, they analyzed daily and weekly patterns as well as several video characteristics such as duration, bit rate, age, ratings, and category. Zink *et al.* [20] also characterized the traffic collected from a university campus. Based on trace-driven simulations, they showed that client-based local caching, P2P-based distribution, and proxy caching can reduce network traffic significantly, allowing faster access to videos.

In common, these studies provide important insights into content popularity and traffic caused by video sharing services. However, they only focused on either on static snapshots of the OSN or on at most a few snapshots [5,9]. Thus, they did not analyze the long-term popularity growth.

Evolution of popularity: A few recent efforts have looked at the time component of content popularity. Regarding videos, Crane and Sornette proposed epidemic models to understand how a popularity burst can be explained in terms of a combination of endogenous user interactions and external events [6]. A similar study was performed by Yang and Leskovec [19] using data from blogging and micro-blogging OSNs. By using a newly proposed time series clustering algorithm, the authors show that content popularity can be grouped in six distinct clusters. Both these studies provide essential tools for time series analysis.

Szabo and Huberman presented a method for predicting popularity of YouTube and Digg[3] content from early measurements of user accesses [17]. More recently, Lerman and Hogg [11] developed stochastic user behavior models to predict popularity based on early user reactions to new content. They improved on predictions based on simple extrapolations from early votes by incorporating aspects of the web site design, validating their approach on Digg.

Another interesting work on popularity evolution in social media was performed by Ratkiewicz *et al.* [16]. By analyzing traffic logs and data from Google Trends[4], the authors investigated how external events, captured by search volume on Google Trends and local browsing (i.e. university/community traffic), affect the popularity of Wikipedia articles. Although this work, and the aforementioned efforts, provide some insights into the long term evolution of content popularity, there is still little knowledge about what kinds of different external events (e.g., being featured on the front page) and system mechanisms (e.g., search) contribute the most to popularity growth. Thus, our analyses and findings, performed separately for YouTube videos with different characteristics, greatly build on previous efforts, shedding more light into the complex task of understanding content popularity on OSNs.

Dynamic OSN analysis: As argued by Willinger *et al.* [18], most previous analyses of OSNs have treated such systems as static. Most of them focus on analyzing structural properties of single snapshots of relationship networks (e.g., friendship network) that emerge in such systems [1,3,15]. Out of the efforts that do exist dealing with dynamic networks, most of the focus on understanding how the structure of the network changes over time [10,12,14]. However, these previous efforts focus on understanding how the structure of the network changes over time. Thus, they do not address how such changes impact system aspects such as content popularity, as we intend to do.

[3] `http://www.digg.com`
[4] `http://trends.google.com`

4. RESEARCH STATEMENT

This thesis aims at addressing the following three research challenges:

RC1 - Patterns of popularity growth: The first research challenge we intend to study, deals with how UGC popularity evolves over time. Analysis on content popularity that deal with only a final number of "hits" have two major limitations: (1) they fail to address the dynamic nature of popularity (e.g. an object may have been very popular in the past only); also, (2) seasonal aspects of popularity are ignored (e.g. bursts or popularity in a seasonal topics). This is an emerging topic of interest in OSNs [6,19].

RC2 - Referrers: Treating individual OSNs as isolated networks on the Internet largely neglects the impact of iterations between the OSN and other websites. For example, the main driving aspect of popularity for a given content may be external to the OSN, such as a blog post or shares in another OSN. Understanding these referrers which lead users to content can be of uttermost importance in determining if UGC is going to be popular in the OSN.

RC3 - Dynamic OSNs: Based on previous arguments [18], we believe that the dynamic nature of the OSN will also have an impact on popularity. Though related, we note that our problem is different from understanding information cascades and epidemics (we refer the reader to Part VI of the Easley and Kleinberg book [7]) in OSNs. We are interested in characterizing how changes in the structural properties of the online social network impact the popularity of UGC shared in the application. Information cascade and epidemics studies mostly analyze information diffusion on static views of OSNs.

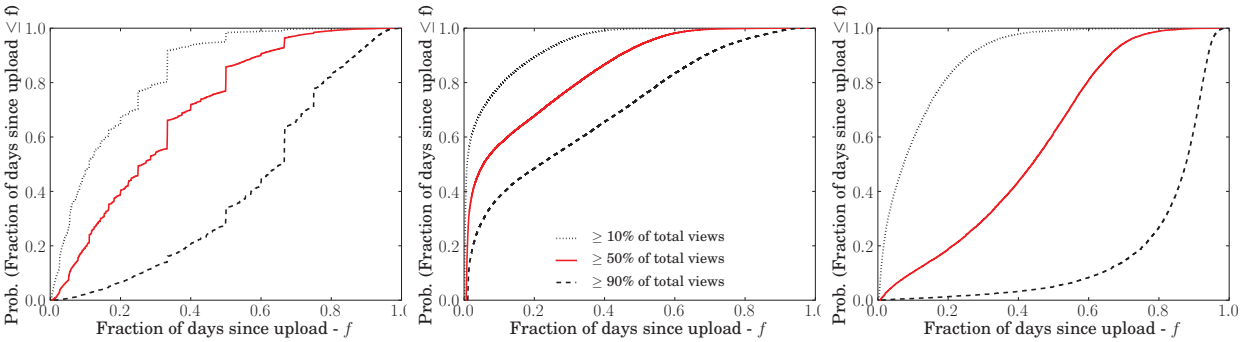
5. PROPOSED METHODOLOGY

The initial step in studying ant of the three challenges above is data collections. Considering *RC 1*, some OSN applications, like YouTube or Delicious [5], provide, on their websites, time series of popularity of different UGC (i.e. videos for YouTube and bookmarks for Delicious). We also note that some notion of popularity can be inferred from user comments on content, a piece of temporal information largely available on many OSNs.

For referrers, some initial insights can be extracted from YouTube which provides the top ten referrers for a video [8]. But in order to collect data on a large scale, we believe that the best option is to monitor traffic to the OSN from a large local network, such as an university campus network [9]. Publicly available data related to previous snapshots of ONSs [6] or network modeling approaches [13] can also be used to address *RC 3*.

After data collection, we intend to exploit the available data to answer the following question: *Can final (or the evolution of) popularity be predicted?* This single research

[5] `http://www.delicious.com`
[6] `http://snap.stanford.edu/data/`



(a) Top (b) YouTomb (c) Random

Figure 1: Cumulative distributions of time until video achieves at least 10%, 50% and 90% of its total views.

question has a major impact on many application level problems of OSN design such as: capacity management and planning; content provision; caching; and, online advertising. We will characterize the data and make use of machine learning techniques (such as regression models and support vector machine based algorithms [2]) in to correlate the three challenges to this question. This approach has been successful in the past on other OSN problems, such as spam detection [3].

Due to the high revenues of online advertising, we believe that this is the most interesting problem which we can apply our results. As a final step in our research, we intend on extending advertising models on OSNs in order to leverage our knowledge on the evolution of content popularity. This step of the work will be done by (1) extending well-known advertisement placement strategies [7] to include our proposed content popularity predictors and (2) simulating these strategies in multiple scenarios of interest.

We next present a brief overview of our current results. We refer the reader to [8] for a detailed description of them.

6. CURRENT STATUS

Using newly provided data by YouTube, we analyzed how the popularity of individual videos evolves since the video's upload time. We also characterized the top ten referrers for each video. Our analysis was performed separately for three YouTube video datasets:

Top: YouTube maintains several top lists (e.g., most viewed and most commented videos), containing one hundred videos each. We crawled these lists gathering 27,212 videos.

YouTomb: Another group of interesting videos is the copyright protected videos. An MIT project, called YouTomb, monitored a large amount of YouTube videos in order to find copyright violations. We used the YouTomb database in order to obtain videos which violated copyrights. We collected a total of 120,862 videos from YouTomb.

Random topics: As basis for comparison, we also studied the popularity growth of a random sample of YouTube videos. We designed a sampling procedure that is based on random topics by submitting random queries to YouTube's search API. We collected 24,484 videos using this strategy.

We now present results on (1) the time interval until a video reaches most of its current popularity (measured according to number of views), which relates to *RC 1*; and (2) the types of referrers that most often lead users to YouTube videos, which are related to *RC 2*.

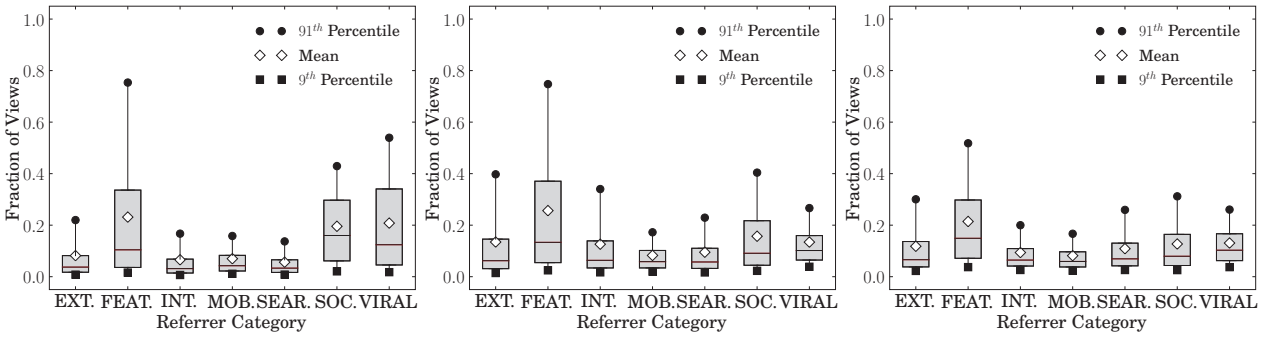
How early does a video get popular? We address this question by plotting, in Figure 1, the cumulative distributions of the amount of time it takes for a video to receive *at least 10%*, *at least 50%* and *at least 90%* of their total views, measured at the time our data was collected. Time is shown normalized by the total time since video was upload, which is here referred to as the video's *lifetime*.

Figure 1 shows that, for half of the videos (y-axis) in the Top, YouTomb and Random datasets, it takes at most 65%, 21% and 87%, respectively, of their total lifetimes (x-axis) until they receive at least 90% of their total views. If we consider at least 50% of their total views, the fractions are 26%, 5% and 43%, respectively, following a similar trend. The same holds for the mark of 10% of the views.

YouTomb videos tend to receive most of their views even earlier possibly because (1) most of them are popular TV shows and music trailers which tend to attract more interest closer to when they are uploaded, and (2) users may seek copyright protected content quicker after upload, before it is removed from YouTube.

Referrer Analysis: We initially grouped referrers into the following categories: External, Featured, Search, Internal, Mobile, Social, and Viral. The *External* category represents websites (often other OSNs and blogs) that have links to YouTube videos. The *Featured* category contains referrers that come from advertises about the video in other YouTube pages or featured videos on top lists and on the front page. On the *Search* category, we group all the referrers from search engines, which comprise only Google services. *Internal* referrers correspond to other YouTube mechanisms, such as the "Related Video" feature, which displays a list of 20 videos that are considered related (according to a YouTube proprietary algorithm) to the watched video. *Mobile* corresponds to all video accesses that come from mobile devices. *Social* referrers consist of accesses coming from the page of the video owner (the channel page) or from users who subscribed to the owner or to some specific topic. Finally, YouTube groups referrers from emails and other sources into a single category, named *Viral*

We analyze the importance of each referrer category, by computing the distributions of the number of views for which each referrer category is responsible, considering only videos that received accesses from the given category. Figures 2(a-c) show box plots containing first, second and third quartiles,



(a) Top (b) YouTomb (c) Random

Figure 2: Distribution of the fraction of views for which each referrer category is responsible.

along with the 9^{th} and 91^{th} percentiles, and the mean, for each referrer category and each video dataset.

From the figure we can see that more than 22% of the views, for the YouTomb videos, come from subscription links in at most for 25% of such videos (3^{rd} distribution quartile). This indicates that users may subscribe to other users that post copyright protected content. The Featured category is a similar case. Moreover, we note that the Social, Featured and Viral categories are responsible for more than 30%, 33% and 34%, respectively, of the views for 25% of the Top videos with referrers from each category (Figure 2-a). Finally, according to Figure 2-c), the Featured category plays a dominant role as source of views to videos in the Random dataset: 25% of the videos that received at least on Featured referrer received at least 30% of their views from such referrers.

7. EXPECTED CONTRIBUTIONS

From the perspective of content popularity analysis in OSNs, this work presented an initial discussion on three different research challenges which will impact the near future of OSN research: (1) understanding evolution of popularity; (2) referrers to content in OSN; and, (3) the impact of network dynamics in popularity. Based on our initial analysis on YouTube, we can conclude that: (1) different types of videos have different popularity growth patterns; and, (2) referrers have a non negligible impact on popularity.

We believe that this research potential to expand the knowledge on OSNs in different aspects. Not only do we intent to provide insights on the three research challenges through thorough characterization content popularity, referrers and dynamic OSNs; we also intent on making use of this knowledge on real world OSN application design issues, such as advertisement placement.

8. REFERENCES

[1] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proc. of the World Wide Web Conference (WWW)*, 2007.

[2] E. Alpaydin. *Introduction to machine learning - 2nd Edition*. The MIT Press, 2010.

[3] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross. Video interactions in online video social networks. *ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP)*, 5(4):1–25, 2009.

[4] S. Boll. Multitube–where web 2.0 and multimedia could meet. *IEEE MultiMedia*, 14(1):9–13, 2007.

[5] M. Cha, H. Kwak, P. Rodriguez, Y-Y. Ahn, and S. Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Network (TON)*, 17(5):1357–1370, 2009.

[6] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proc. of the National Academy of Sciences (PNAS)*, 105(41):15649–15653, 2008.

[7] D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge Univ Pr, 2010.

[8] F. Figueiredo, F. Benevenuto, and J.M. Almeida. The tube over time: characterizing popularity growth of youtube videos. In *Proc. of ACM Int'l Conference on Web Search and Data Mining (WSDM)*, 2011.

[9] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube traffic characterization: A view from the edge. In *Proc. of the ACM Internet Measurement Conference (IMC)*, 2007.

[10] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proc. of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2007.

[11] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *Proc. of the World Wide Web Conference (WWW)*, 2010.

[12] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proc. of the ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining (KDD)*, 2008.

[13] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *The Journal of Machine Learning Research*, 11:985–1042, 2010.

[14] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proc. of the ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining (KDD)*, 2005.

[15] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proc. of the World Wide Web Conference*, 2008.

[16] J. Ratkiewicz, A. Flammini, and F. Menczer. Traffic in social media I: paths through information networks. In *Proc. of the Int'l Symposium on Social Intelligence and Networking*, 2010.

[17] G. Szabo and B. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[18] W. Willinger, R. Rejaie, M. Torkjazi, M. Valafar, and M. Maggioni. Research on online social networks: Time to face the real challenges. *SIGMETRICS Performance Evaluation Review*, 37(3):49–54, 2009.

[19] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proc. of ACM Int'l Conference on Web Search and Data Mining (WSDM)*, 2011.

[20] M. Zink, K. Suh, Y. Gu, and J. Kurose. Watch global, cache local: YouTube network traces at a campus network - measurements and implications. In *Proc. of the IEEE Multimedia Computing and Networking (MMCN)*, 2008.