Video Requests from Online Social Networks: Characterization, Analysis and Generation

Haitao Li, Haiyang Wang, Jiangchuan Liu
School of Computing Science
Simon Fraser University, BC, Canada
Email:{haitaol,hwa17, jcliu}@cs.sfu.ca

Ke Xu
Dept. of Computer Science&Technology
Tsinghua University, Beijing, China
Email: xuke@csnet1.cs.tsinghua.edu.cn

***Abstract*—The deep penetration of Online Social Networks (OSNs) have made them major portals for video content sharing. It is known that a significant portion of the accesses to video sharing sites are now coming from OSN users. Yet the unique features of video sharing over OSNs and their impact remain largely unknown. In this paper, we present a measurement study towards understanding the video requests from OSNs. We closely collaborated with a large-scale Facebook-like OSN to analyze its user access logs spanning over four months. Our measurement reveals a number of distinctive features on the popularity distribution of videos shared over the OSN. In particular, we observe that the OSN amplifies the skewness of video popularity so largely that about 2% most popular videos account for 90% of total views; the video requests distribution also exhibits perfect power-law feature; video popularity evolution shows more dynamics. All these noticeably differ from that of conventional videos, such as YouTube videos. To further understand the characteristics, we model the video viewing and sharing behaviors in OSNs, leading to the development of a practical emulator. It reveals the gap between the sharing rate and the viewing rate, and generates user requests that well capture the video popularity distribution and dynamics as observed in our empirical data.**

I. Introduction

Traditionally, users have discovered videos on the Web by browsing or searching. Recently, word-of-mouth [13] has emerged as a popular way of discovering the videos, particularly over online social network (OSN) sites such as Facebook and Twitter, where users discover video contents following their friends' shares. It has also been a key driving force for the traffic from many video sharing sites (VSSes). A measurement [1] based on YouTube data showed that between April 2009 and March 2010, 25% of views on YouTube come from social sharing. YouTube reported that as of January 2011 more than 500 tweets per minute containing a YouTube link, and over 150 years worth of YouTube video is watched by Facebook users every day [18]. Till June 2012, the numbers have increased to 700 tweets and 500 years. Yet the characteristics of requests from OSNs have not yet been comprehensively measured at large scales, not to mention video requests generation.

To unveil the characteristics of video viewing in OSNs, we closely collaborate with a large-scale Facebook-like OSN in China to analyze its server access logs. Starting from March, 2011, we collected the detailed user video viewing and sharing behaviors over four months. Leveraging the proprietary data, we characterize the user requests from the aspects of video popularity distribution and evolution, unveiling a number of distinctive characteristics compared with the video requests directly from VSSes. In particular, we observe that OSNs amplify the skewness of video popularity so largely that about 2% most popular videos account for 90% of total views (compare to 20%-90% in conventional YouTube statistics [2]). We also observe that the video requests distribution exhibits perfect power-law feature, where in YouTube, it exhibits a power-law waist with a long truncated tail for huge unpopular videos [2].

To further understand the characteristics observed in the empirical analysis, we build an emulator to model the video viewing and sharing behaviors in OSNs. Our emulator generates user requests that well capture the video popularity distribution and dynamics observed in our empirical data. Using this emulator as a tool, we find that although the top popular videos mostly have large *sharing rate* (sharing rate (ShR_i) is defined as the probability viewers will reshare the video i after viewing), videos with high ShR do not definitely gain large user requests. We also confirm that the dynamics of the number of sharers' friends is a major reason for the video popularity dynamics. Our emulator can also be used to synthesize user requests for examining video sharing with assistances from peer-to-peer, content distribution networks, or cloud platforms [16] [17].

The rest of the paper is organized as follows. We present related works in Section II. Section III presents the measurement results on the video popularity distribution. Section IV presents the design, validation, and analysis of our video requests emulator. Finally, we conclude in Section V.

II. Related Work

There are some pioneer data-driven analysis of content propagation in OSNs. Rorigues et al. [13] studied the propagation of URL links posted in Twitter, using large data gathered from Twitter. They presented the distribution of height, width, and size of propagation trees. Sun et al. [15] studied distribution chains and large-scale cascades across Facebook. Scellato et al. [14] focused on the geographic property of social cascades of videos by tracking social cascades of YouTube links over Twitter. Cha et al. [3][4] conducted a large-scale measurement study on Flickr social network. They found that even popular photos spread slowly through the network. While

978-1-4673-5946-7/13/$31.00 ©2013 IEEE

we found that the videos in an OSN spread much faster. This comparison indicates that different kinds of contents propagate in diverse patterns in OSNs. A very recent work [16] studied the propagation-based social-aware replication for social video contents. They found similar power-law video popularity distribution in another large OSN in China. Instead of making a comprehensive measurement and analysis as we do in our paper, they focused on the system optimization based on these new traffic patterns.

Comparing with the characteristics of the videos shared in VSSes can provide us more in-depth understanding of the characteristics of the videos shared in OSNs. There are plenty of measurement works on the VSSes videos either by crawling meta-data their websites [2][6][5] or tracing traffic from a set of network routers/switches [7][20]. Cha et al. [2] presented an in-depth study of the static popularity distribution of videos in two large-scale VSSes, finding that the video popularity shows a power-law waist with a long truncated tail for huge unpopular videos. Figueiredo et al. [6] found that the popularity growth pattern depends on the choice of the video dataset. Crane et al. [5] categorized videos by their popularity evolution patterns into three types: viral videos, quality videos, and junk videos. Gill et al. [7] and Zink et al. [20] both analyzed YouTube video requests from a campus network and observed that the video requests follow a Zipf-like distribution. Our work focuses on similar aspects as pervious works, yet aiming to demonstrate the distinctive characteristics due to the word-of-mouth based sharing mechanism. In particular, we find more skewed popularity distribution, and more complex popularity evolution patterns.

III. Measurement of Video Requests from an OSN

In this section, we first present the measurement results on the video popularity distribution in OSNs. We then study whether these videos always keep the same positions in the distribution and receive corresponding requests over time.

A. Measurement Methodology

To understand the characteristics of the video requests from OSNs, we closely collaborate with an large Facebook-like OSN in China. Like Facebook, users can post video links from VSSes and the video propagation is based on the friend links. Starting from March 24^{th}, 2011, we collect detailed user video viewing and sharing behaviors. Our dataset includes more than 1.1 billion viewing requests over four months. When a user clicks a shared video, an individual record will be sent to the log server. The data format of viewing action is: (Starting Time, Viewer ID, Video URL, Direct Sharer ID, Original Sharer ID). From these trace, we can obtain such statistics as the video popularity distribution, popularity evolution, and inter-arrival times distribution. Moreover, our dataset enables us to analyze the video propagation process in OSNs by tracing viewer-sharer relationships. Tracing the video propagation process helps find the major factors that affect a video's popularity, thus providing inspiration in our model construction.

B. Video Popularity Distribution

The Pareto principle (also known as the 80-20 rule) is widely used to describe the skewness in distributions. Earlier measurement of YouTube shows that 10% of the most popular videos account for 80% of user requests [2]. We expect word-of-mouth based sharing mechanism leads to a less skewed request distribution across the videos in an OSN, since all videos have equal chance to become popular. As shown in Fig. 1, we see a counter-intuitive result that 0.4% videos account for more than 80% of requests; the rest 99.6% of the videos, on the other hand, only account for 20% requests (the x-axis of this figure represents the videos sorted from the most popular videos to the least popular ones, with video ranks being normalized between 0 and 1). We believe that this is because the popular videos will become even more popular since the users are more likely to recommend these videos to their friends. The unpopular videos, however, will fade out very soon in the social communities. An immediate implication from this popularity skewness is that a high hitrate can be achieved, even if only a small set of popular videos are cached.

To further analyze the user requests distribution, we take a closer look at the videos that are initially shared in the same day (March 24^{th}). Since users are generally more interested in newly updated videos, this analysis will avoid the possible bias due to video aging. We count the cumulative requests of those videos after one day, two days, one week and one month respectively, and plot the results in Fig. 2. The popularity of those videos again exhibits such a high skewness that 2% popular videos account for 90% of total requests. We also notice that the skewness increases as the time-window increases, and converges after one week. To further understand the reason for the skewed popularity distribution, we count the frequency of video views in one month, finding that 90% of videos only receive less than 10 requests. These videos were never reshared by viewers since they were introduced to the OSN. It means that these videos ever vanished from the OSN and thus have no chance to be found and requested again. While in VSSes, any videos can always be searched if they are not deleted from the system. The unpopular videos in VSSes can slowly accumulate their views for a long time. Thus, we conclude that the difference in the number of extremely unpopular videos is the direct reason for more skewed video popularity distribution in OSNs than VSSes.

The power-law model [11] has been increasingly used to explain various statistics appearing in the computer science and network systems. To check the power-law pattern for the videos in OSNs, Fig. 3 plots the requests versus video ranks of all videos initially shared on the same day. We find that the plot exhibits perfect power-law (the exponent value is also given in the figure) pattern, and the curves of different days are very similar except for some top videos. While Cha et al. [2] found that the video popularity in YouTube shows a power-law waist, with a long truncated tail for huge unpopular videos and sharp decay for popular videos. They guessed some

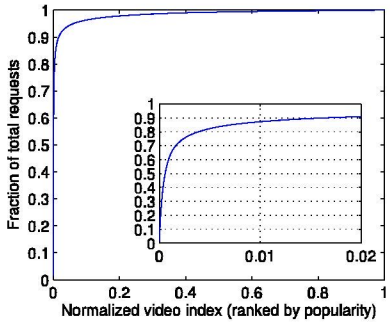

Fig. 1. Skewness of requests across all videos

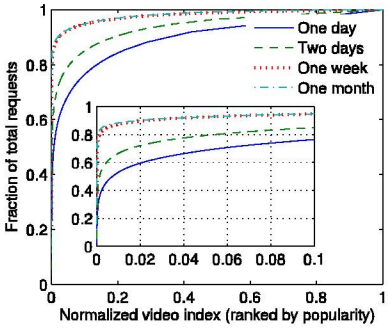

Fig. 2. Videos initially shared in the same day

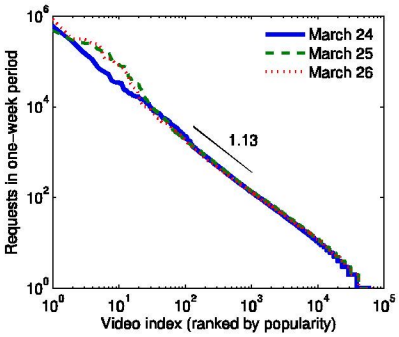

Fig. 3. Requests versus video ranks (log-log)

potential reasons for the truncated tail, but did not provide confident conclusion whether it is the nature shape or because of some design issues. Our later model verifies that it is the nature shape of the power-law video popularity distribution in OSNs. A very recent work [16] also showed the power-law video popularity distribution in another large-scale OSN.

C. Popularity Dynamics

Although the videos show similar popularity distribution along the time, we find that their relative positions in the distribution are highly non-stationary. In other words, some current rarely-requested (or low-ranked) videos may become frequently-requested (top-ranked) videos in the near future. In this subsection, we characterize such dynamics.

For every 500,000 user requests, we snapshot the numbers of added requests across all videos that were initially shared on March 24th. Fig. 4 shows scatter plots for the number of added views received by a video at snapshot 1 and snapshots 2, 3, 4. It also shows the Pearson correlation coefficient (ρ_p) [12] and Spearman's rank correlation coefficient (ρ_s) [10] between the number of added views at different snapshots. With our notion of added views at a snapshot, this figure illustrates the change in viewing rate between two snapshots. Overall, we observe substantial non-stationarity in the popularity of individual videos. Although the added views of two adjacent snapshots shows strong correlation, it is not the case for two non-adjacent snapshots. The correlation declines quickly with the distance of two snapshots. Note that the scatter plots have fewer points for later snapshots owing to the increasing videos that received no views in these snapshots (and hence are not shown on the log-log plots). While in YouTube, prior study found that the early views have relatively high correlation with future views even after one month [2].

IV. Modeling Video Viewing Behaviors in OSNs

In this section, we emulate the users' video viewing behaviors in OSNs. Our emulator is designed to assign a sequence of user requests to a set of videos, and the generated requests should capture the video popularity distribution and dynamics observed in the empirical data. Leveraging this emulator as a tool, we can analyze above measurement results and various factors that impact the video popularity in OSNs. It can also be used to generate synthesize user requests, which are helpful for such related researches as video caching algorithms.

A. Modeling Request Distribution

This paper only focuses on the effect of the word-of-mouth on the dissemination of content. This mechanism is widely adopted by a large number of OSNs (e.g., Facebook, Tittwer, Flicker, and etc.) as the basic information dissemination mechanism. It is also a distinctive feature of OSNs from traditional VSSes. Other mechanisms, such as featuring, links between content, and search results, are undoubtedly at play in some OSNs, but studying their impact requires a richer dataset and is beyond the scope of this paper.

Now we model how the user requests are distributed across videos, in order to capture the video popularity distribution and evolution observed in the empirical data. We denote P_i as the probability that a new request is assigned to the video i. One simple model is distributing requests according to a constant distribution, which is taken in some previous work [8]. This method must assume that the relative popularity of videos maintain stable in a certain time, which is not observed in our empirical data. Another alternative method is using rich-get-richer distribution mechanism [19]. In our case, it is expressed as $P_i = \frac{V_i}{\sum_{j=1}^{M} V_j}$, where M is the number of videos in the system, and V_i is the number of historical views of video i. P_i is proportional to V_i. The most important property for this process is that it generates a distribution following a power law in its tail, as is observed in our empirical dataset. Yet it it can not well reflect the video propagation process in OSNs and hance fail to capture the dynamics of propagation.

Besides viewing history, we try to leverage the video propagation process to provide a more reasonable requests distribution mechanism. Our model assumes that users can only find and view videos shown in the "News Feed" of their homepages. And all these videos are shared by their friends and be pushed to their "News Feed" in a chronological order. Therefore, videos will have more chance to be found and thus be viewed in the future, if they have already been shared by many sharers and at the same time these sharers have plenty of friends. Besides, another two factors are also important: how many of these potential viewers have already watched,

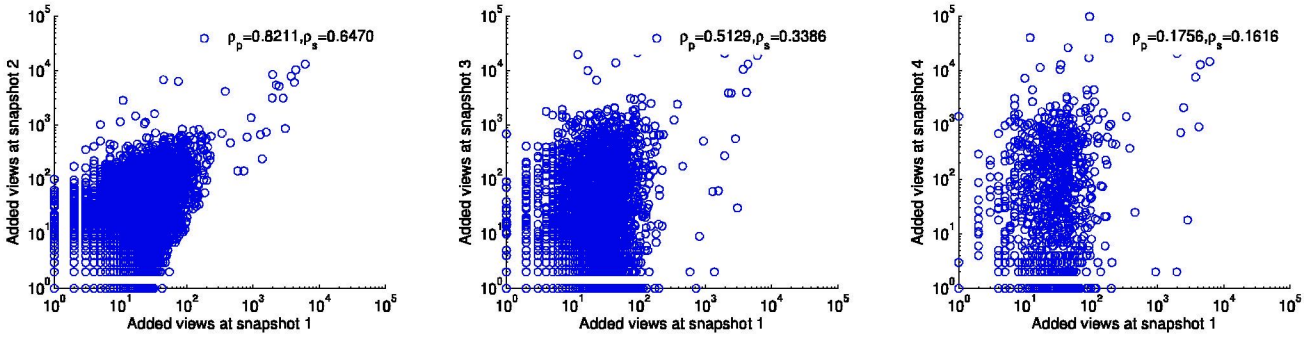

Fig. 4. Scatter plot of the number of added views at snapshot 1 versus snapshots 2, 3, 4

and the probability that users will view the video if it appears in their "News Feed". We define E_i as the expected number of requests for all S_i existing shares of the video i.

$$E_i = \sum_{k=1}^{S_i} (D_i^k * ViR_i) \quad (1)$$

where S_i is the number of sharers of video i until now; D_i^k is out-degree of the k^{th} sharer of the video i; D_i^k indicates how many users the video i can be exposed to if the k^{th} sharer shares it. ViR_i (short for Viewing Rate) is the probability that a viewer will view the video i shared by her/his friend. Note that similar to V_i and S_i, E_i is a variable changing over time. Accordingly, we get the following rich-get-richer equation:

$$P_i = \frac{E_i - V_i}{\sum_{j=1}^{M}(E_j - V_j)} \quad (2)$$

where the value of $E_i - V_i$ reflects the number of expected viewing requests in the future. Larger value of $E_i - V_i$ means more chances to be assigned for the next new request.

B. Emulator

Based on the above model, Algorithm 1 describes an implementation of our emulator for the video viewing and sharing behaviors in an OSN. It introduces a new request to system after each inter-arrival time ($\mathcal{T}$). For each request, the emulator assigns it to the video i according to the P_i defined in Eq. 2. For the chosen video i, the number of its views (V_i) is increased by one. After that, this video should be judged whether to be reshared with the probability ShR_i (short for Sharing Rate of video i). ShR_i . If so, the number of shares (S_i) of this video is increased by one, and the expected views (E_i) of this video is increased by $\mathcal{D} * ViR_i$.

In this emulator, the input parameters include $\mathcal{D}$, video ShR_i, and ViR_i for each video. The emulator distinguishes the attractiveness of videos by assigning different ShR_i and ViR_i to them. Note that it does not distinguish the difference of individual users in the probability of viewing and sharing the same video. The ViR_i and ShR_i are the properties of videos not users. The distribution of $\mathcal{D}$ reflects topological property of the targeted OSN.

Algorithm 1 Emulator of video viewing and sharing behaviors

```
for request = 1 to N do
    generate a inter-arrival time T;
    current time t=t+T;
    a new request arrives, and be assigned to the video i
    with the probability P_i;
    V_i++;
    extract a random variable U, with continuous uniform
    distribution U(0,1);
    if U < ShR_i then
        S_i++;
        extract a random variable D;
        for i=1 to D do
            extract a random variable U;
            if U < ViR_i then
                E_i++;
            end if
        end for
    end if
end for
```

C. Performance Evaluation

We now validate the efficiency of our emulator in reflecting the video popularity distribution and dynamics by inputting the parameters extracted from real-world trace. For the number of videos and requests, we configure the same values (M=63,591 and N=2,905,276) as those in Fig. 3. The distribution of ShR was given in our previous work [9]. Instead of parameterizing ViR and $\mathcal{D}$ separately, the emulator needs only the product of them, which is denoted as BrF (short for Branching Factor). This parameter was also given in our previous work [9].

With the above parameters as the input, we first examine the video popularity distribution of the generated user requests. One key observation from the empirical data about the video popularity distribution is the power-law distribution under the plot of video views versus ranks. Another key observation is that the popularity shows high skewness. As shown in Fig. 5, we can see the simulation result and real-world data are pretty matched. We also count the skewness of the video popularity

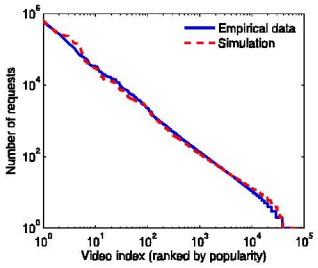

Fig. 5. Popularity distribution

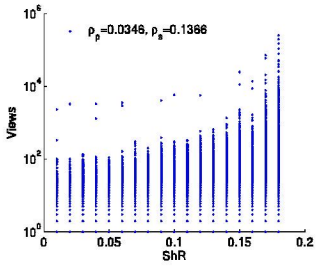

Fig. 6. Impact of ShR

distribution, and the simulation result shows that the top-2% videos account for 85% of the total requests, which is very close to our observation (2%-90%). The most popular videos in our simulation are not as popular as that in the empirical data. In our examined OSN system, a small number (e.g., 120) of popular videos are featured as the most popular videos and are listed in a public page. This behavior can further increase the popularity of the featured videos. We do not include this exogenous factor in our current emulator, considering that it is not a generally case in other systems and also does not affect the overall pattern of user requests.

Then, we examine the popularity dynamics. We calculate the Pearson correlation coefficient (ρ_p) and Spearman's rank correlation coefficient (ρ_s) between the numbers of added views at different snapshots, and shown the results in Table I. Overall, the coefficients are very close for the simulation result and the empirical data. A closer look will find that our emulator produces less dynamics. This is because our simulation simply configures each video with a constant ShR that never changes over time. In fact, the ShR of different videos change over time with diverse patterns, which can also affect the video popularity dynamics. Considering the complexity of ShR evolution pattern yet much less importance to the popularity dynamics, we do not model the evolution of ShR in the current emulator and leave it for our future work.

TABLE I
CORRELATION COEFFICIENTS BETWEEN THE NUMBERS OF ADDED VIEWS AT DIFFERENT SNAPSHOTS (SI VS SJ)

(ρ_p, ρ_s)	S1 vs S2	S1 vs S3	S1 vs S4
simulation	(0.8459,0.7224)	(0.6234,0.3834)	(0.1834,0.1754)
empirical	(0.8211,0.6470)	(0.5129,0.3386)	(0.1756,0.1616)

Based on the verified emulator, we analyze the impact of ShR value to a video's popularity. Fig. 6 shows the scatter plots for ShR and views. On one hand, we find the high ShR does not definitely result in many requests. As shown in this Figure, the correlation coefficients between them are very low. On the other hand, almost all frequently-viewed videos have high ShR. For example, 87.8% videos which gain more than one thousand views have ShR with value 0.17 or 0.18. It indicates the popularity of a video shared in OSN exhibits much randomness and unpredictability, for example owing to the randomness of friends' number of sharers.

V. CONCLUSIONS

In this paper, we studied the characteristics of video requests from OSNs, by analyzing the logs of video viewing actions in a large-scale OSN over serval months. Our measurement unveiled both static and temporal characteristics of video requests from OSNs, highlighting several distinctive features from the requests directly from VSSes. To better understand the characteristics observed in our empirical data, we built an emulator to model video viewing and sharing behaviors in OSNs. Although simple, our emulator well capture the observed characteristics in the empirical data, including the video popularity distribution and dynamics.

ACKNOWLEDGEMENT

This research was supported in part by a Canada NSERC Discovery Grant, an Engage Grant, a MITACS intership grant, and a China NSF international collaboration grant 61120106008.

REFERENCES

[1] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer. Catching a viral video. In *Proc. of ICDM*, 2010.
[2] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. B. Moon. I tube, you tube, everybody tubes: Analyzing the worlds largest user generated content video system. In *Proc. of IMC*, 2007.
[3] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing social cascades in flickr. In *Proc. of WOSN*, 2008.
[4] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proc. of WWW*, 2009.
[5] R. Crane and D. Sornette. Viral, quality, and junk videos on youtube: Separating content from noise in an information-rich environment. In *AAAI Spring Symposium*, 2008.
[6] F. Figueiredo, F. Benevenuto, and J. Almeida. The tube over time: Characterizing popularity growth of youtube videos. In *Proc. of WSDM*, 2011.
[7] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: a view from the edge. In *Proc. of IMC*, 2007.
[8] S. Jin and A. Bestavros. Gismo: A generator of internet streaming media objects and workloads. In *Proc. of SIGMETRICS Performance Evaluation Review*, 2001.
[9] H. Li, J. Liu, K. Xu, and S. Wen. Understanding video propagation in online social networks. In *Proc. of IWQoS*, 2012.
[10] J. S. Maritz. *Distribution-free statistical methods*. Chapman & Hall, 1981.
[11] M. Newman. *Power Laws, Pareto Distributions and Zipfs Law*. 2004.
[12] J. L. Rodgers and W. A. Nicewander. *Thirteen ways to look at the correlation coefficient*. The American Statistician, 1988.
[13] 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.
[14] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft. Track globally, deliver locally: Improving content delivery networks by tracking geographic social cascades. In *Proc. of WWW*, 2011.
[15] E. Sun, I. Rosenn, C. Marlow, and T. Lento. Gesundheit! modeling contagion through facebook news feed. In *Proc. of ICWSM*, 2009.
[16] Z. Wang, L. Sun, X. Chen, W. Zhu, J. Liu, M. Chen, and S. Yang. Propagation-based social-aware replication for social video contents. In *Proc. of ACM Multimedia*, 2012.
[17] Z. Wang, L. Sun, C. Wu, and S. Yang. Guiding internet-scale video service deployment using microblog-based prediction. In *Proc. of Infocom*, 2012.
[18] YouTube. http://www.youtube.com/t/press_statistics.
[19] G. Yule. A mathematical theory of evolution based on the conclusions of dr. j. c. willis. *Philosophical Transactions of the Royal Society of London*, 1925.
[20] M. Zink, K. Suhb, Y. Gu, and J. Kurosea. Characteristics of youtube network traffic at a campus network - measurements, models, and implications. *Computer Networks*, 2009.