Understanding the External Links of Video Sharing Sites: Measurement and Analysis

Kunfeng Lai, *Student Member, IEEE* and Dan Wang, *Member, IEEE*

***Abstract*—Recently, many video sharing sites provide external links so that their video or audio contents can be embedded into external web sites. For example, users can copy the embedded URLs of the videos of YouTube and post the URL links on their own blogs. Clearly, the purpose of such function is to increase the distribution of the videos and the associated advertisement. Does this function fulfill its purpose and what is the quantification? In this paper, we provide a comprehensive measurement study and analysis on these external links to answer these two questions. With the traces collected from two major video sharing sites, YouTube and Youku of China, we show that the external links have various impacts on the popularity of the video sharing sites. More specifically, for videos that have been uploaded for eight months in Youku, around 15% of views can come from external links. Some contents are densely linked. For example, comedy videos can attract more than 800 external links on average. We also study the relationship between the external links and the internal links. We show that there are correlations; for example, if a video is popular itself, it is likely to have a large number of external links. Another observation we find is that the external links usually have a higher impact on Youku than that of YouTube. We conjecture that it is more likely that the external links have higher impact for a regional site than a worldwide site.**

I. Introduction

These years have witnessed large explosion of the popularity of the online UGC (user generated content) sites. In these sites, people are not only the information consumers, but they can also actively upload contents of their own. Different UGC sites have different emphasis. For example, Facebook is built as a general online community, Flickr is best known as a photo sharing site, and twitter is unique in its short message distributions. Among the UGC sites, this paper will focus more specifically on video sharing sites, which are best represented by YouTube [36] and Youku [37].

A common belief of the success of the UGC sites is that the information generated by users can be distributed much faster through the UGC sites. For example, the video sharing sites provide numerous functionalities to expedite video distribution. There are the related video links (See Fig. 2) which arrange videos by similar topics. There are mechanisms inside video sharing sites to organize users and videos together. Many previous works have studied these mechanisms in details and we refer interested readers to [13][20][32].

To further popularize the video distribution, video sharing sites introduce external links. An example of the external link is shown in Fig. 1. We can see that for each video in YouTube, an embedded link is provided. The user can copy and paste this embedded link into anywhere such as their personal webpages, blogs, or even forums. When people watch the videos outside the video sharing sites, traffic and click counts go through YouTube. Clearly, the external links allow YouTube videos to be embedded in non-YouTube sites to attract views. This can further accelerate video distribution.

Fig. 1. An example of an external link of YouTube.

We see that these external links are very different from those functions and features that arrange the internal contents, such as the videos and users. To be more specific, we define the internal interaction as user-to-user, user-to-video, and video-to-video relationship inside the video sharing sites. We define the external interaction as the referencing of the videos outside the video sharing sites, such as hyperlinks to the videos.

Fig. 2. An example of an internal link, the related video link (R-link).

In this paper, we are interested in these external links. Compared with past studies on the interaction between users and videos within the video sharing sites, we are the first to concentrate on external links to videos of these sites. We have the following contributions in this paper: 1) we proposed to study the external links of the video sharing sites and we tried to quantify its impact. We believe this adds to the knowledge base, and could be useful for future comparison; 2) we showed that the impact from external links is non-trivial and we also found substantial differences on the impact of the external links on YouTube and Youku; 3) we conducted measurements

K. Lai and D. Wang are with Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. E-mail: {cskflai, csdwang}@comp.polyu.edu.hk. A preliminary version of this paper appeared in Proc. ACM NOSSDAV'10, Amsterdam, the Netherland, June, 2010.

on both external links and some important internal links and we studied their correlations; 4) we published the data sets of external links we collected from YouTube and Youku for possible follow-up studies. The data sets can be found in [34].

The remaining part of our paper is organized as follows. We present the related work in Section II. Section III discusses the background and measurement methodology. The impact of the external links on the videos is given in Section IV. In Section V we study the correlation between internal links and external links. In Section VI, we focus on the evolution of the external links by studying the external links of the videos of different uploading time. Finally, we conclude our paper in Section VII.

To assist reading, we have a summary subsection on the main observations at the end of Section IV, V and VI.

II. Related Work

Nowadays, there are increasing interests on the UGC (User Generated Content) sites. In these sites, the users not only consume the contents, but also freely create the contents their own. Due to their large popularity and success, there have been many studies on different types of UGC sites. For example, Facebook, a general online community, is studied in [28], and Flickr, a photo sharing site is studied in [27].

In this paper, we study video sharing sites. The successful examples of this type of UGC site are YouTube, which enjoys world-wide popularity and Youku, the biggest video sharing site in China [33].

We are not the first to study the video sharing sites. Among past studies, measurement methodologies have been proposed. For example, in [3] guidelines are provided to do sampling in the video sharing sites and a framework is also proposed to study the popularity dynamics of user-generated videos. An important conclusion is that using key-word search videos as the seeds, videos would be biased towards the popular videos, while using recently uploaded videos, there would have no bias. The study in [31] provides a method to count the total number of YouTube videos by leveraging the video ids. A globally distributed active measurement platform is established in [1] for YouTube video delivery system.

The analysis on the video sharing sites widely spans to user-to-user, user-to-video, and video-to-video relationship. For example, the video popularity distribution of YouTube is analyzed in [4], where the long tail of video popularity distribution is analyzed. More aspects of YouTube are analyzed in [7], such as video life cycles, user viewing behaviors, and the small world phenomenon. In [13], the traffic of YouTube in campus is characterized. It shows that there is strong correlation between videos viewed (watched) on consecutive days. This work also demonstrates that caching can improve user experiences, reduce bandwidth consumption and lower the burden of YouTube. The recommendation system in YouTube is studied in [32]. It shows 30% of the video total views come from related video links. The user behavior in the video sharing sites is also widely studied. For example, the study in [9] focuses on the YouTube video uploaders, and demonstrates positive reinforcement between online social behavior and uploading behavior. In [11], it is shown that user access patterns are similar in different locations and their access devices (either PCs or mobile phones): they usually use default video resolution and player configuration. The work in [2] shows an interesting result that online video consumption appears geographic locality of interest. The video-to-video relationship is analyzed in [29], and it shows that more than half of the YouTube videos contain re-mixed video segments, and some particularly popular videos are correlated with virus.

There are suggestions to improve YouTube. NetTube is developed using a peer-to-peer structure for YouTube [7]. An algorithm using geographic information to improve multimedia content delivery in YouTube is suggested in [23]. The study in [30] suggests an improved semi-supervised training method for classifying YouTube videos by using video labels and co-watch relationship (videos watched in one session).

There are also studies that compare YouTube with other UGC sites. For example, the video popularity distribution of four different video sharing sites is characterized in [21]. It shows that the life time video popularity have relevance with caching size. A comparison of different UGC sites (including YouTube) can be found in [20] and observations of the free scale, small world and strong connected cores, are drawn. The study in [14] compares the user sessions between the video sharing sites and the traditional web sites. It concludes that the YouTube users have larger data traffic and longer think time.

We can see that existing studies on video sharing sites all focus on internal interactions; that is, the content-to-content, user-to-content, and user-to-user relationship inside the UGC sites. In this paper, we focus on understanding the characteristics of the external links. We have some measurement and analysis of the related videos of YouTube and Youku; though the emphasis is on comparison with the results of external links. We have two preliminary works [16][17]. In [16], we present a short study on YouTube only. This paper substantially extends [17] to the domains such as the correlation between external links and internal links in all video age groups.

III. Background and Measurement Methodology

A. Background and Motivation

Currently, there is rapid growth of the popularity of user generated content web sites. One key feature of these sites is that the users are not only the information consumers, but also actively upload contents of their own. One notable class of UGC sites is for video sharing, represented by YouTube and Youku. These video sharing sites have attracted a great number of studies in the recent years. These studies, however, focus on user-to-user, user-to-video or video-to-video relationship within these video sites. To distribute the content videos more widely and to attract more users, these video sites provide external links for videos. Users can easily obtain an embedded link of a video and paste the link to any web pages in other web sites, such as forums, or their blogs. In this paper, we define the *internal links* as those maintaining a relationship within the web sites. These links include the user-to-video, user-to-user, video-to-video relationship. We define the *external links* as the links to the videos that are embedded in other web sites.

These external links are important for improving the popularity of the videos; however, there is no rigid study to quantify the effectiveness of these external links. Therefore, we would like to know 1) the impact of the external links on videos, e.g., how many views are contributed by external links; 2) the relationship of external links and internal links; their differences, interactions and correlations. Such curiosities motivate this paper.

B. Measurement Methodology

Our experimental data sets come from two video sharing sites, YouTube [36] and Youku [37]. YouTube is one of the largest video sharing sites in the world and at the time of this paper is being written, it accepts 1.886 billion views [33] every day. Youku is the most popular video site in China [33], with views of 40.9 million per day [33].

The necessary data for our study are 1) a large number of randomly collected videos, and 2) the external links of these videos. We first built a crawler to sample a large base of videos. In principle, we start our data sampling from seed videos and follow their related videos in the crawling. We follow [10][3], where it is shown that if those recently-uploaded videos do not link popular videos as their only related videos, the data collection will not be biased by choosing the seed videos from recently-uploaded videos.

More specifically, for YouTube, we started our sampling on Mar. 24th, 2009 using the “recent” videos as seeds. We recursively crawled all the related videos for seven days until Mar. 31st, 2009. In total, we have collected 1.24×10^6 videos from YouTube. For Youku, unlike YouTube, there is no category of “recent” videos. Therefore, we used all videos in the main pages, which were uploaded within one day, on July 8th, 2009 as seeds, and recursively collected the related videos for five days. In total, we have collected 1.43×10^6 videos. We admit that for these videos, they might have more views.

To collect the external links, we used a universal Java Script engine provided by Google [38]. This engine can parse the Java Script codes from video pages, so as to track the external link information maintained by YouTube and Youku internally. With this engine, we could get the URLs of the external links as well as the number of views of each external link. However, from YouTube and Youku, we can only have the information of “top” external links, which are calculated based on the number of views contributed to the videos since the videos were uploaded. YouTube maintains the top-5 external links of each video. We have not found a method that can collect the information of all the external links of videos. Intrinsically, if YouTube does not provide an interface to release such information, unless one can explore the entire Web, it is unlikely that all external links can be collected. In our study, we use the top-5 external links for YouTube videos. For Youku, we used similar method. Youku provides more information, and we obtained the total number of external links for each video, the URLs and the total number of views of the top-20 external links.

We admit that collecting information only from top external links affects the accuracy of the study. Our argument is that, on one hand, part of our studies, e.g., comparison of the total views from these external links in different video age groups, is less affected by the total external links. On the other hand, for the Youku trace, we have the total number of external links, and the number of views of each top-20 external link. Thus, we have a strong basis to analyze the remaining part of the views of external links. We did find that the average views from external links as a function of the rank of external links in a large sampling space fit the power law function with an r-square over 99.8% (see Section IV.C). Therefore, we are more confident that our observations of this study are close to reality. In the remaining part of our paper, we simply say external links of YouTube and external links of Youku, which should be understood that we denote the top-5 external links for YouTube and the top-20 external links for Youku, given that there is no ambiguity.

Besides the external links, we also collected the information of internal links (we focus on the related video links) for a comparison study. For such data collection we adopt the same strategies as [4]. We also started on Mar. 24th, 2009 and July 8th, 2009 and carried out a data collection of 7 days and 5 days for YouTube and Youku respectively. The data collection is a breadth first search, following the related video links.

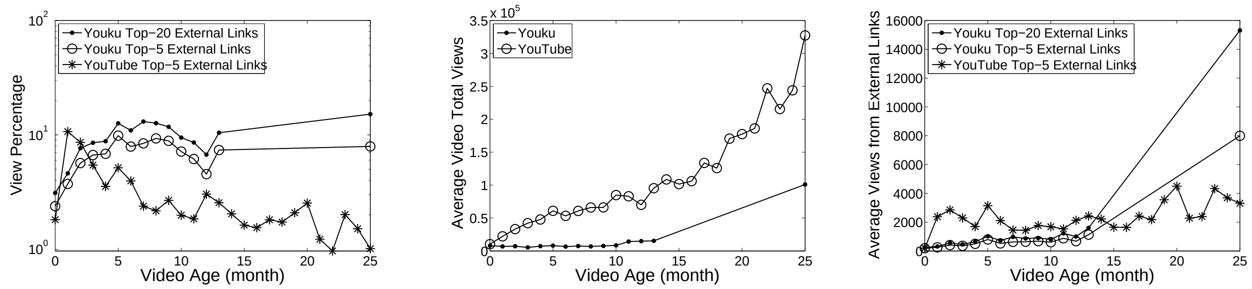
IV. The Contribution of the External Links

A. Overall Contribution of External Links

We first show the impact of the external views on the videos in Fig. 3. We classify the videos according to their *ages*, i.e., the total duration since they have been uploaded to the video sharing sites. Note that YouTube provides the upload date for each video, whereas Youku provides a rougher estimation of how many days or months or years a video has been uploaded. For example, the videos uploaded 13 months or 14 months ago in Youku will both be labeled as ‘uploaded one year ago’. As such, the points of 13-month and 25-month in our figures for Youku stand for the videos uploaded one year and two years ago. Note that our results are not affected as the points in our figures are the average (not accumulative) number of views.

In Fig. 3 (a), we show the percentage of the views that come from the top external links. We see that for the videos in YouTube with an age of two months, 10% of the views come from the top-5 external links. For videos with an older age, the percentage of the views from external links gradually drops to around 2%. For Youku, the impact of external links is much higher. For most of the videos, more than 8% of views come outside the video sharing site. For videos with an age of 24 months, views from external links can contribute as many as 15%. Even considering the top-5 external links of Youku, they contribute about 6% - 9% of total views, which is still more significant than YouTube.

To explain the situation more clearly, we show the specific number of the video total views and views from the top external links as a function of video ages. In Fig.3 (b) we show the total views, averaged per video, for different video age groups (this includes both views from internal links and external links). The total views increase steadily for both YouTube and Youku as video ages increase. It is also clear that



(a) External views vs. video age (b) Total views vs. video age (c) External views vs. video age

Fig. 3. The contribution of external links on the popularity of videos on videos sharing sites.

YouTube attracts much more views than Youku. This is not surprising as YouTube is more popular world-wide. In Fig.3 (c), we show the total views from external links (averaged per video). We see that for YouTube, the total external views are comparatively stable among all video age groups whereas for Youku, the total external view increases. In addition, though the external views of YouTube are still greater than that of Youku, the differences are not as big as the total views. This explains why the percentage of external views of Youku is more significant than that of YouTube in Fig. 3 (a).

Such differences of the impact of the external links on YouTube and Youku do not conform to our expectation. We consider a possible explanation can be as follows. YouTube is a video sharing site of world-wide popularity. As such, the external links are widely spread to external websites all over the world. These external websites may not be world-wide popular, however. Thus, the external links cannot obtain world-wide popularity and have less impact (in terms of percentage). Youku, on the contrary, has popularity within China only. The external links are also on the Chinese-based websites and can have China-wide popularity. As such, the comparative impact of external links on Youku is much higher than on YouTube. Based on our current data, we are unable to verify this. As a first work on external links, we confine ourselves to the fundamental problems such as the correctness of the data collection, and the understanding of the basic characteristics of external links, as will be presented in the remaining part of the paper. We will leave such questions to our future work.

B. The Number of External Links

t (month)	$\alpha(t)$	*s(t)*	x_{min}	KS statistic
1	6.508×10^3	1.450	20	0.018
7	1.490×10^4	1.519	80	0.0052
12	2.293×10^5	1.806	120	0.0016

TABLE I
THE PARAMETERS OF POWER LAW FITS FOR FIG. 4

Fig. 4 plots the number of videos as a function of the number of external links in a log-log scale for different age groups. Since YouTube cannot provide the total number of the external links for each video, we only study Youku in this figure. We can see clearly that a small portion of videos enjoy the majority of the external links. With the methodology introduced in [8],

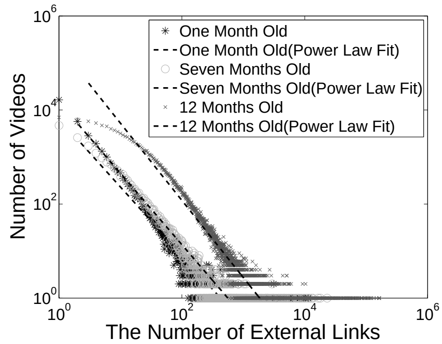


Fig. 4. The distribution of the number of external link of videos in different age groups.

we therefore use Maximum Likelihood Estimation (MLE) to fit power law distribution $f(x) = \alpha \times (\frac{x}{x_{min}})^{-s}$ (Table I shows the specific parameters α, s and x_{min} for each age group) to the sampled data. We find that in Youku the power law distribution fairly matches the external link distribution. This is especially true for the videos with a large number of external links, i.e., the tails of the distribution. We find that the 'tails' follow power law with Kolmogorov-Smirnov statistics (KS statistics) respectively 1.8%, 0.5% and 0.2%. This is not entirely surprising. Notice that the power law behavior has been observed in various properties of the video sharing sites. For example, it is shown that the distribution of the number of video views in YouTube also fits the power law [4][6].

We classify the videos into three age groups (one month, seven months, and 12 months). We study the percentage of videos with more than ten external links. We see that different age groups show clear trend: for the videos in age group of one month, 13.4% of videos have more than ten external links; for the videos in age group of seven months and 12 months, 33.0% and 64.8% of videos have more than ten external links. This shows that in Youku older videos, on average, may get a larger number of external links.

C. The Views from External Links

In this subsection, we study the views contributed by external links. We still use the Youku data set for analysis. Fig. 5 shows the views of the external links as a function of the ranks of the external links in a log-log scale. Note that the *rank* is from one to 20 as we have the views of the top-20 external links. We plot the data for the videos with an age of one month, seven months and 12 months in Fig. 5. We found the data approximately follow power law function where there are deviations at the tails. The power law means that most of the external views come from the high rank external links.

To further analyze how the views of external links decay as the rank of external links increases, in Fig. 5, we use the least square method to fit the original data with power law function $\mathcal{V}_1(t) = a_1(t) \times r_1^{-p_1(t)}$ where t is the age of the video group, r_1 is the rank of external links, $p_1(t)$ is an exponential factor and $a_1(t)$ is an adjustment factor (Table II shows $p_1(t), a_1(t)$). Since the tails have deviation, we also plot a power law function with a deviation term ϵ, $\mathcal{V}_2(t) = a_2(t) \times r_2^{-p_2(t)} + \epsilon(t)$ (Table III shows $p_2(t), a_2(t)$, and $\epsilon(t)$). We can see that 1) the power law functions fit the original data well (as it is shown in Table II and Table III, all the $\mathcal{V}_1$ and $\mathcal{V}_2$ fits are with an r-square greater than 0.99); 2) the views from high rank (larger than the rank of ten) of external links are slightly larger than the power law fit $\mathcal{V}_1$ but slightly smaller than the fit $\mathcal{V}_2$.

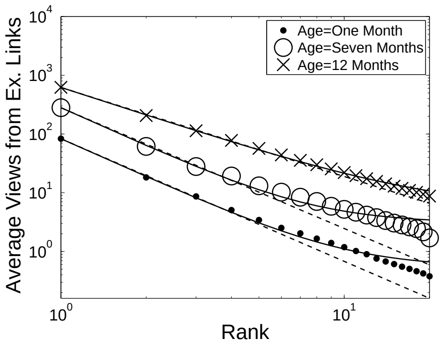


Fig. 5. External views as a function of rank (all videos). The dashed lines shows the power law fit lines $\mathcal{V}_1(t) = a_1(t) \times r^{-p_1(t)}$, and the solid line shows the fits $\mathcal{V}_2(t) = a_2(t) \times r^{-p_2(t)} + \epsilon(t)$

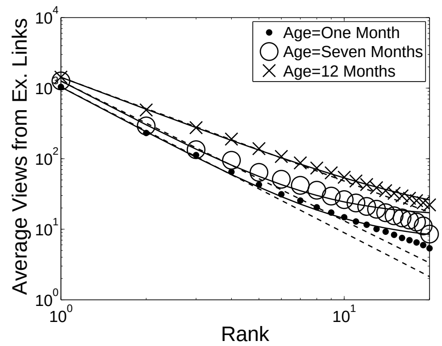


Fig. 6. External views as a function of rank (number of links $\geq$ 20). The dashed lines shows the power law fit lines $\mathcal{V}_1(t) = a_1(t) \times r^{-p_1(t)}$, and the solid line shows the fits $\mathcal{V}_2(t) = a_2(t) \times r^{-p_2(t)} + \epsilon(t)$

Note that we only have the number of views for the top-20 external links. Since we have seen a power law fit, we conjecture that the views of top-20 external links are representative for the views of all external links. As a result, we estimate the total views from all external links. This will make the conclusions of this paper more grounded.

t (month)	$a_1(t)$	$p_1(t)$	r-square
1	82.98	2.087	0.9998
7	279.9	2.055	0.9999
12	619.3	1.515	0.9999

TABLE II
THE PARAMETERS OF POWER LAW FITS FOR FIG. 5

t (month)	$a_2(t)$	$p_2(t)$	$\epsilon(t)$	$r - square$
1	82.55	2.165	0.54	0.9998
7	277.4	2.089	3.06	0.9999
12	615.9	1.580	5.31	0.9999

TABLE III
THE PARAMETERS OF POWER LAW FITS WITH A CONSTANT DEVIATION TERMS FOR FIG. 5

To estimate the total views from all external links, we divide the videos into two groups: 1) the videos with more than 20 external links and 2) the videos with less than 20 external links. For Group 2, Youku provides the total views from the top-20 links of each video, or the views of the top-k links if the video only has k external links and $k \leq 20$. Let the total views from external links of Group 2 be V_2^*. For Group 1, we can further divide it into two parts: a) the total views from the top-20 external links and b) the total views from other (non-top-20) external links. Youku provides a). Let this be V_{20}. Thus, our primary target is to estimate part b) of Group 1.

t (month)	$a_1(t)$	$p_1(t)$	r-square
1	1037	2.369	0.9978
7	1259	1.981	0.9985
12	1424	1.471	1.0

TABLE IV
THE PARAMETERS OF POWER LAW FITS FOR FIG. 6

In Fig. 6 we plot the total views of the external links as a function of the ranks of the external links in a log-log scale for the videos with more than 20 external links. We plot the data for the videos with an age of one month, seven months and 12 months. We see that the number of views also is close to the power law function. Again, to fit the power law function, we also let $\mathcal{V}_1(t)$ and $\mathcal{V}_2(t)$ be the total number of views for videos with age t months for the fits. We plot power law function

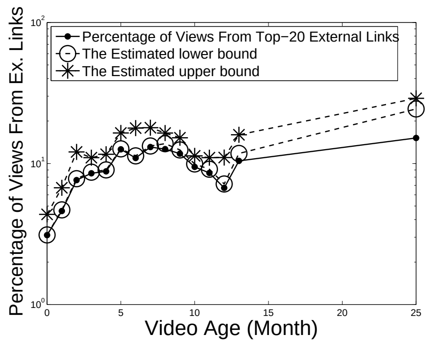
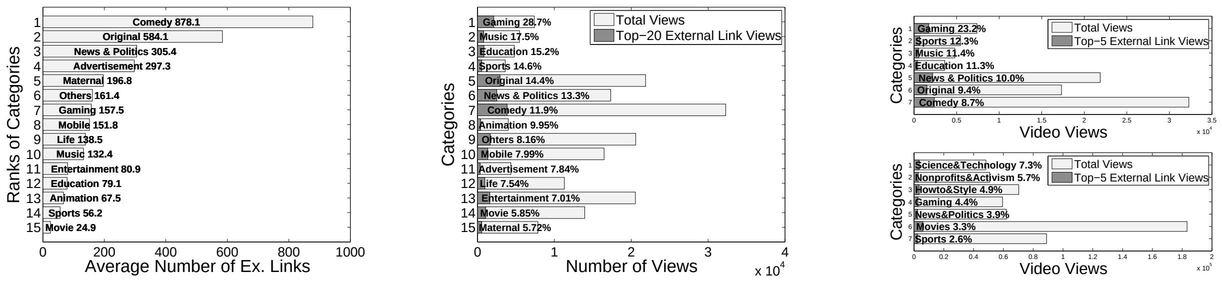


Fig. 7. The estimated views from external links of the Youku videos.



(a) The average number of external links per video

(b) Average views from top-20 external links per video (Youku)

(c) Average views from top-5 external links per video

Fig. 8. The contribution of external links on different categories of videos from both YouTube and Youku

t (month)	$a_2(t)$	$p_2(t)$	$\epsilon(t)$	$r - square$
1	1031.6	2.1429	6.77	0.9998
7	1248	1.9203	14.95	0.9998
12	1416	1.529	11.49	1.0

TABLE V
THE PARAMETERS OF POWER LAW FITS WITH A CONSTANT DEVIATION TERMS FOR FIG. 6

$\mathcal{V}_1(t) = a_1(t) \times r_1^{-p_1(t)}$ and the power law function with deviation $\mathcal{V}_2(t) = a_2(t) \times r_2^{-p_2(t)} + \epsilon(t)$ for different age groups in Fig. 6 (Table IV and Table V show $p_1(t), a_1(t), p_2(t), a_2(t)$, and $\epsilon(t)$, as well as the r-squares of the fits). We also see that the number of views from higher rank of external links is slightly larger than $\mathcal{V}_1$ but smaller than the fit $\mathcal{V}_2$.

Let $V_1^*(t)$ be the total views for part b) of Group 1, and let $\mathcal{N}(t)$ be the average number of external links of the videos aged t months, then

$$\sum_{i=21}^{\mathcal{N}(t)} (a_1(t) \times i^{-p_1(t)}) \leq V_1^*(t) \leq \sum_{i=21}^{\mathcal{N}(t)} (a_2(t) \times i^{-p_2(t)} + \epsilon) \quad (1)$$

Therefore, the estimated external link total views are $V_{20} + V_1^*(t) + V_2^*$. We plot both lower bound and upper bound of percentage of the external link total views contributing to the video total views in Fig. 7. We also plot the percentage of views from top-20 external links. Clearly, both the lower bound and the upper bound of the external link total views are greater than the views of the top-20 external links (which is only a section of the video total views). However, we can see that the views of the top-20 external links are very close to the estimated external link total views. Most of the time, the views of the top-20 external links contribute to 90% of the total views from all external links. As such, we conclude that the total views from the top-20 external links are representative for our study.

D. The External Links from Different Video Categories

We next study from the point of view of the videos. We plot the number of external links of different video categories in Youku in Fig. 8 (a). In this figure, the categories are ranked by the average external links on each video. We see that the number of external links of videos is substantial. For example, for an average comedy video, there can be as many as 878.1 external links. Looking into the details of our log file, we see that many comedy videos are linked by a substantial number of different users in their blogs, usually by copying and referring of others' blogs. Some videos are linked in a great many pages in web forums. This actually suggests that external links can greatly increase the popularity of the videos.

In Fig. 8 (b), we select 15 categories in Youku, which have the highest percentage of external views. We plot the average total views of each category (in green) and the average views of the top-20 external links (in red). In the figure, "Comedy" attracts the largest number of external views, on average 3847.5 per video. This is not surprising as "Comedy" also attracts the largest number views (32286.2), representing the popularity of comedy videos in general. We also see that "Gaming" attracts 7391.4 views in total and 2125.0 views from external links, where the external views share the highest percentage. This suggests that as compared with other video categories in Youku, more percentage of views in the "gaming" category come outside Youku.

To compare Youku and YouTube, we show the views of different categories based on their respective views of the top-5 external links (see Fig. 8 (c)). In general, YouTube attracts an order of magnitude more views than Youku, but the percentage of external views is much smaller. This conforms to the observation in Fig. 3. Another observation is that the categories most viewed by external links are substantially different in YouTube and Youku. For example, they share in common only 3 out of 7 categories, namely "Gaming", "Sports", and "News & Politics". This might show different tastes of the users throughout individual regions. In addition, the more success in Youku in extending its impact of the external links could suggest that there may be also potential for YouTube to increase its external views.

E. Summary

We summarize our major observations in this section as follows: 1) the sheer number of external views and the external links are substantial for both YouTube and Youku. The external views/links have contributed greatly to Youku while it still remains small to YouTube; 2) Most of the external links are linked to a small number of videos, i.e., the number of external links conforms to power law distribution; it fits especially well for the videos with large number of external links; 3) The number of external views also conforms to power law; it fits better for old-age videos. Though we cannot obtain the total views from all external links, with the observation of power law, we can deduct that the views from top-20 external links are representative enough for videos; 4) Different video categories attract different percentages of external views. In some categories, e.g., "Gaming" in Youku, almost 30% of views are contributed by external links.

V. External Links vs. Internal Links

Since external links contribute to the video popularity, we analyze the factors that affect the number of external links. We study the relationship between internal links and external links which respectively represent the internal interactions and the external interactions. According to Alexa [33], a user spends an average of 22 minutes on YouTube and 6.7 minutes on Youku every day. As a result, we infer that many users would watch multiple videos in the video sharing sites. For these users, there are many ways to view multiple videos in the video sharing sites, and one of them is to follow the related video list (See Fig. 2). We call a video the *parent video* for the videos in its related video link list. We specifically focus on the relationship of the external links, the related video links (We call them the *R-links* thereafter) and the total views from parent videos (We call them *parent views* thereafter). In addition, we also study the relationship between the external links and other factor, such as the total views of videos.

A. Internal Parameters and External Links

Fig. 9 presents the relationship of total views of a video with two internal factors, namely the number of R-links and parent views, in both YouTube and Youku. We plot in Fig. 9 (a) and (b) the number of total views as a function of the number of R-links of Youku and YouTube respectively. Here, we hardly see any impact of the R-links on the improvement of the total views in Youku, but we see that clear correlation exists for YouTube. This shows that more R-links lead to an increase of the views of YouTube. We plot in Fig. 9 (c) and (d) the average views of a video as a function of the views from the parent videos for Youku and YouTube respectively. We see that in general, the parent views have a positive impact on the video views for both Youku and YouTube, and the parent views of YouTube show an even stronger impact.

Fig. 10 depicts the relationship between the number of external links and the three internal factors, namely the total views, the views from parent videos and the number of R-links. All the results in Fig. 10 are from Youku data set. Fig. 10 (a) shows a clear relationship between the total views and the number of external links. Especially when the number of views is over 100, we can see this relationship is almost linear. Fig. 10 (b) shows the views from the parent videos has weaker correlation with the number of external links. The number of external links scatters as the views of parent videos grows. In Fig. 10 (c), we see there is even weaker relationship between the number of R-links and the number of external links.

$Corr$	views	Ex. links	R-links	parent views
views	1	0.506	-0.018	0.22
Ex. links	0.506	1	-0.029	0.20
R-links	-0.018	-0.029	1	0.23
parent views	0.22	0.20	0.23	1

TABLE VI
THE CORRELATION COEFFICIENT OF PARAMETERS IN YOUKU

$Corr$	views	R-links	parent views
views	1	0.49	0.77
R-links	0.49	1	0.60
parent views	0.77	0.60	1

TABLE VII
THE CORRELATION COEFFICIENT OF PARAMETERS IN YOUTUBE

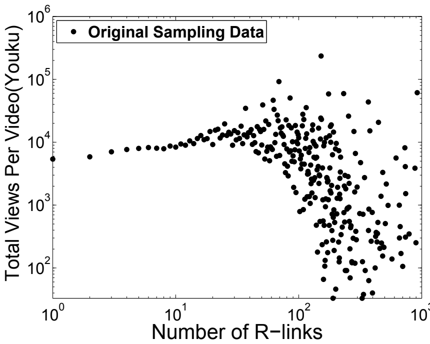
B. Analysis of the Correlation

We next conduct analysis on the correlation coefficients[1] between the number of external links and internal factors for both Youku and YouTube (see Table VI and Table VII). In Table VI we find that the total views and the number of external links of Youku are most correlated with a correlation coefficient of 0.506. The number of R-links hardly affects the number of external links and the total views. The views from parent videos weakly correlated with all the factors. This conforms to the general intuition that the popularity of the video itself will directly impact on the number of external links (and vice versa). Also, we see that in Youku, the views of the parent videos have a more moderate impact and a larger related video links can hardly have an impact on the external links.

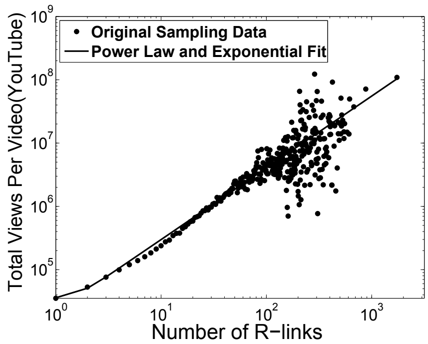
Compared with Youku, the connections of internal links in YouTube are much tighter. In details, we can see that the relationship among the number of R-links, the parent views and the total views are high. The correlations are 0.77 between video total views and parent views, 0.60 between the number of R-links and parent views, and 0.49 between the number of R-links and video total views respectively.

A possible explanation of the tighter correlation in YouTube is the stability of the related video list; that is, if the related videos change more frequently, it will affect the correlation between R-links and other factors. To verify this, we randomly chose 20,000 videos from the crawled video data set of both Youku and YouTube. We crawled/searched these videos from Youku and YouTube again on Nov. 28th 2011. We see that as compared to the 20,000 videos collected on Jul. 8th, 2009, 5376 Youku videos have been deleted and as compared to the 20,000 videos crawled on May 24th, 2009, 387 YouTube videos have been deleted. As the videos in YouTube are significantly more stable than Youku, we infer that there is less change in the related videos of YouTube too. Therefore, YouTube has a tighter correlation of internal links than Youku.

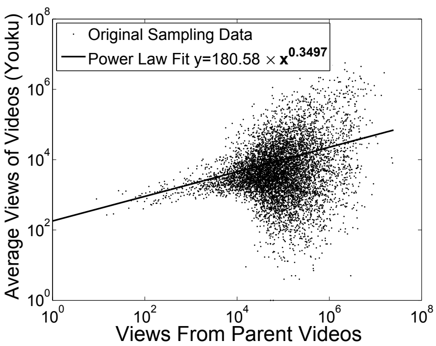
[1]Here, a correlation coefficient of 1 indicates that the two parameters are linearly correlated, i.e., one parameter will increase (or decrease) linearly with the other parameter. A correlation coefficient of -1 indicates that one parameter will increase linearly as the other parameter decreases.



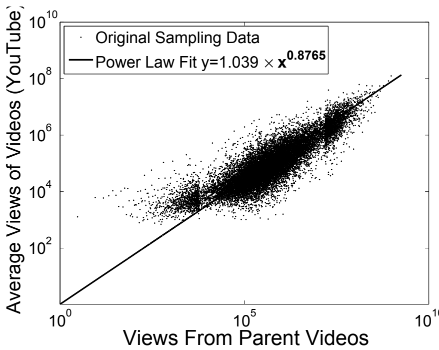
(a) The number of R-links as a function of total views (Youku)



(b) The number of R-links as a function of total views (YouTube)

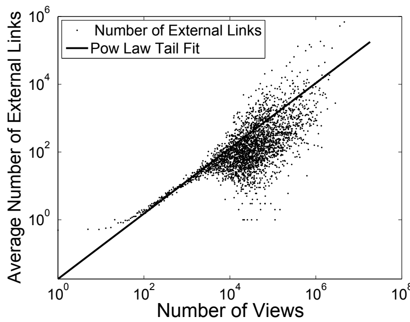


(c) Views of parent videos as a function of total views (Youku)

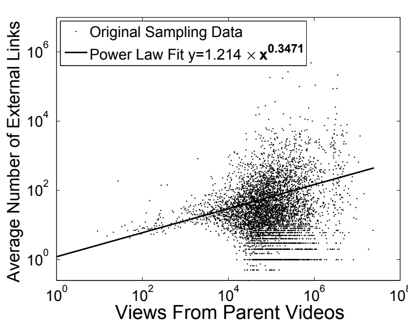


(d) Views of parent videos as a function of total views (YouTube)

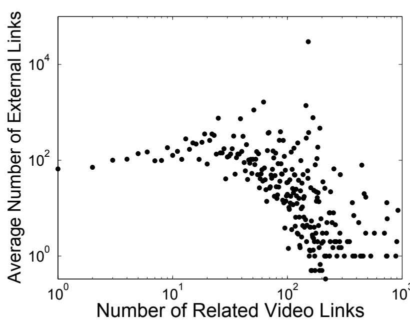
Fig. 9. The relationship between the video views, the number of external links, the views from parent videos, and the number of R-links



(a) Video views as a function of the number of ex. links



(b) Views of parent videos as a function of number of external links



(c) The number of R-links as a function of the number of ex. links

Fig. 10. The relationship between the number of external links and various internal factors (such as the video total views, the views from parent videos and the number of R-links)

C. Summary

We summarize our major observations in this subsection as follows: 1) the number of external links is mostly and more directly affected by the total views of the videos; 2) the number of external links can be affected indirectly by such internal factors, such as parent views and the number of R-links, since these factors can increase the views of videos; 3) the internal factors in YouTube have a stronger correlation than that of Youku. This may be because the Youku videos are less stable as there is a higher deletion rate of the videos.

VI. External Links on Videos in Different Age Groups

To further understand the impact of external links, as well as the correlation of external links with various internal factors, in this section, we investigate the characteristics of external links in different video age groups. Since we did not trace specific external links or trace the external links for specific videos, we group the videos according to different ages. Our study is then on the characteristics of external links on younger videos and older videos. We believe this provides a macro view of the evolution of the external links on videos.

A. The External Links on Videos of Different Age Group

We first compare the percentage of videos received external links in YouTube with that in Youku. We only focus on the videos that have five or more external links.

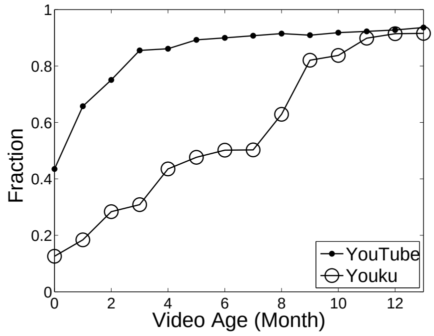


Fig. 11. The percentage of the videos with more than five external links.

From Fig. 11 we can see that videos in YouTube are more linked by external links than videos in Youku. For example, it is observed that for YouTube, 90% of the five-month-old videos have at least five external links. Looking from another angle, we can say that 93.7% of the videos are with at least five external links after 15 months. On the contrary, in Youku, only about 50% of the five months old videos have at least five external links. Nevertheless, the percentage rises eventually and from the data we collected, we can see that after a video that is one year old, more than 90% of the chance (both for YouTube and Youku) it will have more than five external links. The differences between YouTube and Youku may be accounted as follows. First, compared with Youku, YouTube is a world-wide video sharing site and the videos in YouTube have a larger audience base. Therefore, the YouTube videos gain popularity more quickly. This has been proved in Fig. 3 (b) where we see the YouTube videos have a larger number of average views in each age group. Second, as we have shown in Section V.B, the number of external links for each video has a positive relationship with the video total views. Therefore, as the YouTube videos gain popularity (being viewed) more quickly, the YouTube videos also get external links faster than Youku videos.

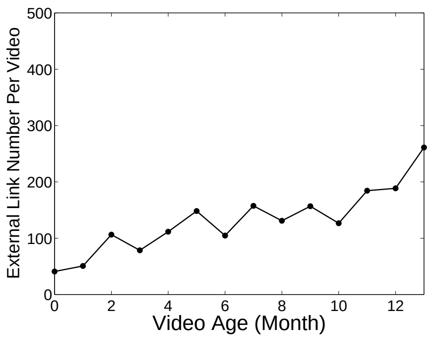


Fig. 12. The average number of external links for each video of different video ages, Youku.

We then study the average number of external links for each video of different video ages. As YouTube can only provide the top-5 external links for each video, we thus focus on Youku data. The results are in Fig. 12. We can see that the average number of external links is increasing. This is not surprising that the total views of videos are increasing with video ages (as it is shown in Fig. 3 (b)), and there is a positive relationship between the video total views and the number of external links (as it is shown in Section V.B). As well, the average views from external links for each video of different video ages are also increasing, as shown in Fig. 13.

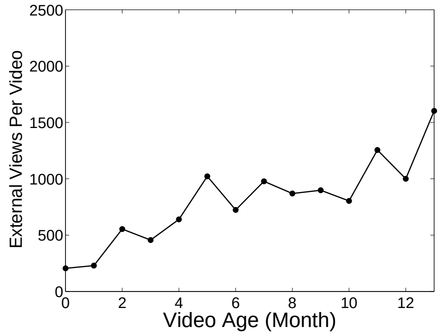


Fig. 13. The average views from external links for each video of different video ages, Youku.

B. The Correlation between External links and Video Total Views in Different Video Age Groups

We study the correlation of the number of external links and the video total views according to different age groups. From Section V, we see that there is a positive correlation between the number of external links and the total views (including both external views and internal views).

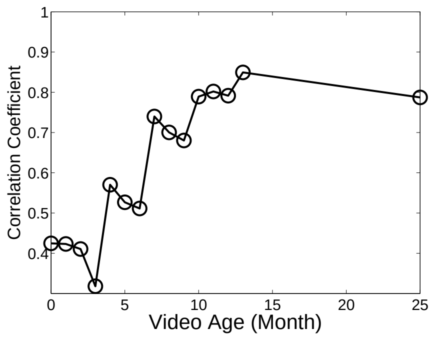


Fig. 14. The correlation of external link number and video total views with the video ages

From Fig. 14, we see that the correlation becomes stronger for the videos in older age groups. The correlation coefficient rises from 0.42 for the video group of one month old to about 0.85 for the video group of one year old. This indicates that the relationship between the number of external links and the video total views is strengthened with time. We conjecture that there is positive impact from both sides, i.e., 1) the number of external links and external views increase, contributing to the video total views, and 2) the video total views (and thus the popularity) increase, contributing to the increase of the external links.

C. The Correlation Coefficients in Different Video Age Groups

In Fig. 15, we show the correlation coefficient among external links, video total views, parent views, and the number of R-links in YouTube and Youku.

Fig. 15 (a) shows the correlation coefficient of the number of external links with the parent views and the number of R-links in Youku. We can see that the correlation coefficient between the number of external links and the parent views fluctuates sometimes but remains above 22% after the first month. However, the correlation coefficient between the number of external links and R-links remains zero. This conforms to the observations in Section V.B.

Fig. 15 (b) shows the correlation of video total views with parent views and the number of R-links in Youku in different video age groups. Not surprisingly, we can see that in Youku the number of R-links do not contribute to video total views in any video age group. However, similar with the relationship between the number of external links and the video total views, we can see the correlation between the video total views and parent views is also strengthened as videos get older, from 0% to about 40%. As the correlation between the number of external links and video total views are strengthened with the video age, and the relationship between video total views and parent views is also positive, we infer the following: for Youku, if a video can gain more parent views, the total views and the number of external links of the video may be increased after a period of time.

Fig. 15 (c) plots the correlation of video total views with parent views and the number of R-links in YouTube in different video age groups as a comparison. Here, we can see that these two correlation coefficients maintain stable and high. As compared to Youku, we can still see that the correlation of internal links in YouTube is significantly larger. This also conforms to the results in Section V.B.

D. Summary

As a summary, as the videos get older, the number of external links of each video is increasing as well as the number of external views of each video. This indicates the increase of the number of external links is not restricted in certain ages of videos but for all videos. The correlation between the number of external links and the video total views is strengthened with the video ages. As we can also see a positive correlation coefficient between video total views and parent views, we conjecture that if a video can obtain more parent views, the number of external links may be increased.

VII. Conclusion

In this paper, we studied in detail an important aspect of video sharing sites: the external links. The external links provide a unique way for the video sharing sites to accelerate the distribution of the videos. We observed that the external links can play a non-trivial role both in terms of the number of external links on a video, and the number of views contributed to the video. We also observed that the external links have quite different impacts on YouTube and Youku. We studied the external links for different video categories. We also discussed the correlations of the external links and the internal related video links. We showed that the number of internal related video links have less impact on the external links than the total views of the video. We also study the characteristics of external links in different video age groups. We see that videos are possible to get external links and external views in all age groups. We believe that our work can provide the foundation for the video sharing sites to make more targeted advertisement, customized user development, etc.

As a first work on the external interactions of video sharing sites, we concentrate on some fundamental problems, such as how the data of external links can be collected, whether the data collection on top external links can provide a good approximation for the overall picture, and some basic aspects of the external links. There are problems yet to be answered. Especially, we are interested in more detailed analysis of the different impacts of external links on Youku and YouTube.

VIII. Acknowledgements

Dan Wang's work was supported by Hong Kong PolyU/APJ19, A-PK95, A-PL23, and A-PL86.

References

[1] V. Adhikari, S. Jain, Y. Chen, and Z. Zhang, "Vivisecting YouTube: An Active Measurement Study", In *Proc. IEEE Infocom'12*, Orlando, Florida, USA, March 25 - 30, 2012.

[2] A. Brodersen, S. Scellato, and M. Wattenhofer, "YouTube Around the World: Geographic Popularity of Videos", In *Proc. ACM WWW'11*, Lyon, France, April 16 - 20, 2011.

[3] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "Characterizing and Modelling Popularity of User-generated Videos", In *Performance Evaluation*, pp. 1037-1055, 68, October 2011.

[4] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S, Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System", In *Proc. ACM IMC'07,* San Diego, California, USA, October 24 - 26, 2007.

[5] W. Chen, J. Chu, and J. Luan, "Collaborative Filtering for Orkut Communities: Discovery of User Latent Behavior," In *Proc. ACM WWW'09,* Madrid, Spain, April 20 - 24, 2009.

[6] X. Cheng, C. Dale, and J. Liu, "Statistics and Social Network of YouTube Videos", In *Proc. IEEE IWQoS'08,* Enschede, The Netherlands, May 28 - 30, 2008.

[7] X. Cheng, and J. Liu, "NetTube: Exploring Social Networks for Peer-to-Peer Short Video Sharing", In *Proc. IEEE Infocom'09,* Rio de Janeiro, Brazil, April 19 - 25, 2009.

[8] A. Clauset, C. Shalizi, and M. Newman, "Power-Law Distributions in Empirical Data", In *SIAM Review,* Vol. 51, No. 4, 2009.

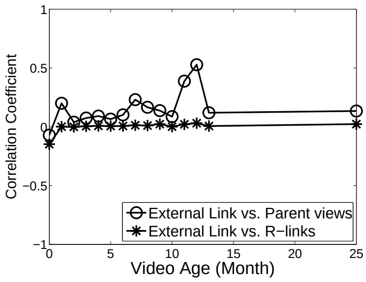
[9] Y. Ding, Y. Du, Y. Hu, Z. Liu, and L. Wang, "Broadcast Yourself: Understanding YouTube Uploaders", In *Proc. ACM IMC'11,* Berlin, Germany, November 2 - 4, 2011.

[10] F. Figueiredo, F. Benevenuto, and J. Almeida, "The Tube over Time: Characterizing Popularity Growth of YouTube Videos", In *Proc. ACM WSDM'11,* Hong Kong, China, February 9 - 12, 2011.

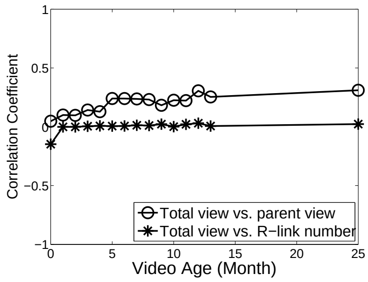
[11] A. Finamore, M. Mellia, M. Munafo, R. Torres, and S. Rao, "YouTube Everywhere: Impact of Device and Infrastructure Synergies on User Experience", In *Proc. ACM IMC'11,* Berlin, Germany, November 2 - 4, 2011.

[12] A. Fuxman, P. Tsaparas, K. Achan, and R. Agrawal, "Using the Wisdom of the Crowds for Keyword Generation", In *Proc. ACM WWW'08,* Beijing, China, April 21 - 25, 2008.

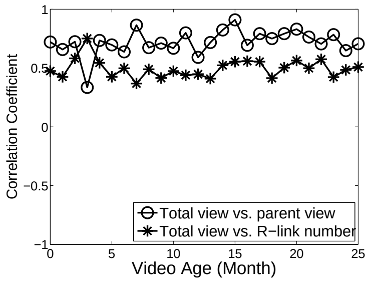
[13] P. Gill, M. Arlittz, Z. Li, and A. Mahanti, "YouTube Traffic Characterization: A View From the Edge", In *Proc. ACM IMC'07,* San Diego, California, USA, October 24 - 26, 2007.



(a) The corr. of external links with other factors (Youku)



(b) The corr. of video total views with other factors (Youku)



(c) The corr. of video total views with other factors (YouTube)

Fig. 15. The evolution of correlation of external links and other internal factors with time

[14] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Characterizing YouTube user sessions", In *Proc. SPIE Multimedia Computing and Networking Conference MMCN'08*, San Jose, CA, January 20 - 31, 2008.

[15] J. Jiang, C. Wilson, X. Wang, P. Huang, and W. Sha, "Understanding Latent Interactions in Online Social Networks", In *Proc. ACM IMC'10,* Melbourne, Australia, November 1 - 3, 2010.

[16] K. Lai and D. Wang, "A Measurement Study of External Links of YouTube", In *Proc. IEEE Globecom'09,* Hawaii, USA, November 30 - December 4, 2009.

[17] K. Lai and D. Wang, "Towards Understanding the External Links on Video Sharing Sites: Measurement and Analysis", in *Proc. ACM NOSSDAV'10,* Amsterdam, Netherland, June 2 - 4, 2010.

[18] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney, "Statistical Properties of Community Structure in Large Social and Information Networks", In *Proc. ACM WWW'08,* Beijing, China, April 21 - 25, 2008.

[19] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting Positive and Negative Links in Online Social Networks", In *Proc. ACM WWW'10,* Raleigh, USA, April 26 - 30, 2010.

[20] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks", In *Proc. ACM IMC'07,* San Diego, California, USA, October 24 - 26, 2007.

[21] S. Mitra, M. Agrawal, A. Yadav, N. Carlsson, D. Eager, and A. Mahanti, "Characterizing Web-based Video Sharing Workload", In ACM Transactions on the Web, Vol.5, No. 2, May 2011.

[22] B. Ribeiro, and D. Towsley, "Estimating and Sampling Graphs with Multidimensional Random Walks", In *Proc. ACM IMC'10,* Melbourne, Australia, November 1 - 3, 2010.

[23] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft, "Track Globally, Deliver Locally: Improving Content Delivery Networks by Tracking Geographic Social Cascades", In *Proc. ACM WWW'11,* Hyderabad, India, March 28 - April 1, 2011.

[24] G. Szabo and B. Huberman, "Predicting opularity of online content", In *Communications of the ACM*, Vol. 53, Issue 8, August, 2010.

[25] J. Tang, M. Musolesi, and C. Mascolo, "Temporal Distance Metrics for Social Network Analysis", In *Proc. ACM SIGCOMM WOSN'09,* Bacelona, Spain, August 17, 2009.

[26] M. Torkjazi, R. Rejaie, and W. Willinger, "Hot Today, Gone Tomorrow: On the Migration of MySpace Users", In *Proc. ACM SIGCOMM WOSN'09,* Bacelona, Spain, August 17, 2009.

[27] M. Valafar, R. Rejaie, and W. Willinger, "Beyond Friendship Graphs: A Study of User Interactions in Flickr", In *Proc. ACM SIGCOMM WOSN'09,* Bacelona, Spain, August 17th, 2009.

[28] B. Viswanath, A. Mislove, and M. Cha, "On the Evolution of User Interaction in Facebook Bimal", In *Proc. ACM SIGCOMM WOSN'09,* Bacelona, Spain, August 17, 2009.

[29] L. Xie, A. Natsev, J. Kender, M. Hill, and J. Smith, "Visual Memes in Social Media: Tracking Real-World News in YouTube Videos", In *Proc. ACM MM'11,* Scottsdale, Arizona, USA, November 28 - December 1, 2011.

[30] J. Zhang, Y. Song, and T. Leung, "Improving Video Classification via YouTube Video Co-Watch Data", In *Proc. ACM SBNMA'11,* Scottsdale, Arizona, USA, December 1, 2011.

[31] J. Zhou, Y. Li, V. Adhikari, and Z. Zhang, "Counting YouTube Videos via Random Prefix Sampling", In *Proc. ACM IMC'11,* Berlin, Germany, November 2 - 4, 2011.

[32] R. Zhou, S. Khemmarat, and L. Gao, "The Impact of YouTube Recommendation System on Video Views", In *Proc. ACM IMC'10,* Melbourne, Australia, November 1 - 3, 2010.

[33] "Alexa", http://www.Alexa.com.

[34] "The Data Sets of External Links from YouTube and Youku", http://www4.comp.polyu.edu.hk/~cskflai/external-link.html.

[35] "Kaixin is not happy: suffering from the loss of 60% traffic in one year" http://news.xinhuanet.com/it/2011-08/04/c_121812027.htm

[36] "YouTube", http://www.youtube.com.

[37] "Youku", http://www.youku.com.

[38] "V8 - V8 JavaScript Engine", http://code.google.com/p/v8/ .

Kunfeng Lai (S11) received his B.S. degree from University of Electronic Science and Technology of China, Chengdu, China, 2005, and the M.S. degree from the same university in 2008; all in computer science and technology. He is currently a Ph.D student of the Hong Kong Polytechnic University. His research interests include online social network measurement and cloud computing. He is a student member of the IEEE.

Dan Wang (S05-M07) received the B.Sc degree from Peking University, Beijing, China, in 2000, the M.Sc degree from Case Western Reserve University, Cleveland, Ohio, USA, in 2004, and the Ph. D. degree from Simon Fraser University, Burnaby, B.C., Canada, in 2007; all in computer science. He is currently an assistant professor at the Department of Computing, The Hong Kong Polytechnic University. His research interests include wireless sensor networks, Internet routing and applications. He is a member of the IEEE.