The Impact of YouTube Recommendation System on Video Views

Renjie Zhou[†], Samamon Khemmarat[‡], Lixin Gao[‡]

[†] College of Computer Science and Technology
Harbin Engineering University, Harbin, China
renjie_zhou@hrbeu.edu.cn

[‡] Department of Electrical and Computer Engineering
University of Massachusetts, Amherst, USA
{khemmarat,lgao}@ecs.umass.edu

ABSTRACT

Hosting a collection of millions of videos, YouTube offers several features to help users discover the videos of their interest. For example, YouTube provides video search, related video recommendation and front page highlight. The understanding of how these features drive video views is useful for creating a strategy to drive video popularity. In this paper, we perform a measurement study on data sets crawled from YouTube and find that the related video recommendation, which recommends the videos that are related to the video a user is watching, is one of the most important view sources of videos. Despite the fact that the YouTube video search is the number one source of views in aggregation, the related video recommendation is the main source of views for the majority of the videos on YouTube. Furthermore, our results reveal that there is a strong correlation between the view count of a video and the average view count of its top referrer videos. This implies that a video has a higher chance to become popular when it is placed on the related video recommendation lists of popular videos. We also find that the click through rate from a video to its related videos is high and the position of a video in a related video list plays a critical role in the click through rate. Finally, our evaluation of the impact of the related video recommendation system on the diversity of video views indicates that the current recommendation system helps to increase the diversity of video views in aggregation.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General; H.3.5 [**Information Storage and Retrieval**]: Online Information Services

General Terms

Measurement, Human Factors

IMC'10, November 1–3, 2010, Melbourne, Australia.

Keywords

Video Sharing Site, YouTube, Recommendation System, View Sources, View Diversity

1. INTRODUCTION

YouTube has been one of the most successful user-generated video sharing sites since its establishment in early 2005. It is estimated that there are over 100 million videos on YouTube [9], which makes the exploration of the desired videos a daunting task. In order to help users find interesting videos from a huge number of videos, YouTube provides several features such as search engine, front page highlight, and related videos recommendation.

YouTube can obtain the data on how users use these features, which can be useful for improving its service. However, the understanding of how video views are driven by these features is also helpful and interesting to outsiders such as researchers, advertisers, and content providers. For researchers, this is a case study of how contents in a large repository are discovered and the evidence for the importance of content discovery tools. For content providers and advertisers, this is useful for planning strategically to increase their videos' popularity and predicting the effectiveness of advertising. More recently, YouTube started to let video content providers be partners to cash in on the videos posted by sharing ad revenue and charging rental fees to viewers [2]. This underscores the need to understand how one can drive the views of a video.

In this paper, we perform a measurement study on data sets of hundreds of thousands of videos crawled from YouTube website. We study how videos are discovered by users, what are the major sources that drive the views of a video, and how well related video views are correlated. We summarize our findings as follows.

- The related video recommendation is one of the most important view sources of a video. Despite the fact that the YouTube video search is the number one source of views in aggregation, the related video recommendation is the main source of views for the majority of the videos on YouTube. In particular, for a large portion of videos with lukewarm popularity, the major source of their views is from users clicking related videos.
- There is a strong correlation between the view count of a video and the average view count of its top referrer videos. This means that if the top referrer videos are popular, then the video is also popular. This implies a video has a higher chance to become popular when

it is placed on the related video recommendation lists of popular videos. Furthermore, the position that a video is placed on the related video list plays a critical role in the click through rate of the video.

- We evaluate the impact of the video recommendation system on the diversity of video views. In contrast of recent results on the recommendation on book/CD sales [6], we find that YouTube recommendation provides more diversity on video views in aggregation than that without the recommendation. This means that YouTube recommendation helps viewers discover videos of their interest rather than popular videos only.

The rest of the paper is organized as follows. In section 2, we describe the data sets used in this study. In Section 3, we study the view sources for overall videos as well as individual videos. In Section 4, we investigate how the related video recommendation system affects video views. The evaluation of the impact of the related video recommendation system on the diversity of video views is studied in Section 5. Related work is described in Section 6 and finally, Section 7 concludes the paper.

2. DATA DESCRIPTION

Our study is based on the data sets crawled from YouTube. In this section, we first describe three elements of data provided by YouTube. We then describe how we collected the data sets.

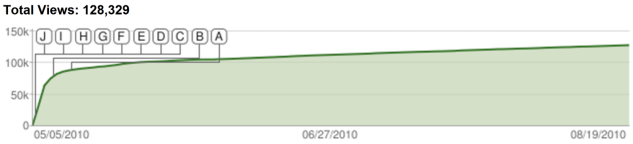
2.1 Data Source

On YouTube, a video is viewed on a page named *watch page*, which not only shows the video itself, but also includes valuable data about the video. We focus on three elements of the data. The first element is *video metadata*, which includes the basic information about the video, such as title, upload time, and total view count. The second element is a *related video list*, which contains the related videos recommended by YouTube recommendation system. The third element is *view statistics & data*, which includes the particular video's view count sequence over time, top ten view sources, and the date of the first referral and the number of views from each source, as shown in Figure 1. By investigating the view statistics & data, we find that YouTube classifies the sources into 14 categories, such as YouTube Search, Related Video, and Mobile Device. Each of them indicates a specific kind of source, except for the Other/Viral category, which includes all the view sources that do not fall into other categories. For the category of Related Video, it also shows the video that leads users to the current video. We refer to this video as a *referrer video* of the current video.

2.2 Data Collection

In this section, we describe how we collected the two data sets in this study. We start by describing how the data can be obtained. The video metadata and related video list can be retrieved through HTML scraping as well as YouTube Data API, while the view statistics & data can only be retrieved through HTML because it is not supported by the API.

We used two different video sampling methods to obtain our data sets to minimize the bias from data sampling. Our results in the next section show that the trends of the results from the two data sets are consistent with each other despite



Links

	Date	Link	Views
A	May 13, 2010	First referral from related video - MILEY CYRUS CAN'T BE TAMED SPOOF SHOOT!	4,600
B	May 10, 2010	First referral from related video - Did it all by myself	1,995
C	May 07, 2010	First referral from related video - Ke$ha BTS music video "Your love is my drug"	12,408
D	May 06, 2010	First referral from a subscriber module	34,601
E	May 06, 2010	First view from a mobile device	13,044
F	May 06, 2010	First referral from YouTube - Homepage	12,429
G	May 06, 2010	First embedded on - ijustine.com	3,010
H	May 06, 2010	First referral from YouTube search - zhu zhu pets	2,685
I	May 06, 2010	First referral from YouTube - /user/ijustine	2,430
J	May 06, 2010	First referral from YouTube - /videos	1,492

Figure 1: Snapshot of view statistics & data.

the different methods of video sampling. In the following, we describe the crawling process for each data set. For D1, we selected the videos by capturing and parsing YouTube video requests at a university network gateway, and the three elements of data associated with each video was collected through crawling. For D2, we retrieved 400 featured videos via API as the initial set, crawled their associated data, and then crawled the data for their related videos. We did this recursively in a breadth first search manner for three levels. Additionally, we crawled the metadata of referrer videos for both D1 and D2 and the related video lists of referrer videos for D2.

The data we collected and the approaches we used in the collection are shown in Table 1, and the amount of data we collected is shown in Table 2.

Collected Data	D1		D2	
	Crawl	Method	Crawl	Method
Metadata	Y	API	Y	API
Related Videos	Y	API	Y	API
Statistics & Data	Y	HTML	Y	HTML
Related Videos of Referrers	N	N/A	Y	API

Table 1: Collection method for each data set.

Data set	D1	D2
Start Date	28-Jan-10	17-Mar-10
Duration	21 days	14 days
# Videos with Metadata	498,233	202,428
# Videos with Related Videos	154,363	202,428
# Videos with S&D	111,351	55,280
# Referrer Videos	348,059	133,114

Table 2: General statistics of the data sets.

3. SOURCES OF VIDEO VIEWS

YouTube videos are accessed in a variety of ways, such as through Google search, Facebook, mobile device, and through the features provided on YouTube itself. In this section, we study which view source is the most frequently used and which one is the major contributor of views for the majority of the videos by investigating video view statistics & data. As mentioned in Section 2.1, view statistics & data

contains the top ten view sources and the number of views from each source. We refer to the total views driven by the top ten sources as the *tracked views* of a video. It is necessary to verify first whether the tracked views are representative.

3.1 Representativeness of Tracked Views

To justify that the tracked views are representative, we use the percentage of tracked views from the total views of a video, and the correlation between them as the criteria. We calculate the percentage of tracked views for each video in D1 and D2, and the average is 56.7% and 63.1%, respectively. This means that, on average, the tracked views are the majority of the total views. Further, the Cumulative Distribution Function (CDF) of the percentage of tracked views is shown in Figure 2. From the figure, there are around 70% of videos whose tracked views are larger than 50% of the total views, which means that for the majority of videos, more than half of their views are recorded in view statistics & data.

In addition, we compute the Pearson's correlation coefficient between tracked views and total views to measure the linear correlation between them. The definition of Pearson's correlation coefficient is given by

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

where $\rho_{X,Y}$ is the correlation coefficient between total views X and tracked views Y with the expected value μ_X and μ_Y, and the standard deviation σ_X and σ_Y, respectively. The Pearson's correlation coefficients for D1 and D2 are 0.81 and 0.84, respectively. This indicates a strong linear correlation between tracked views and video views in both data sets. The correlation can be seen clearly in Figure 3, in which we plot the number of tracked views and total views for each video.

The high percentage of tracked views and the strong linear correlation between tracked views and total views lead to our conclusion that the tracked views can represent the total video views well. We believe that the view pattern explored by analyzing the tracked views is a good approximation of the overall view pattern on YouTube.

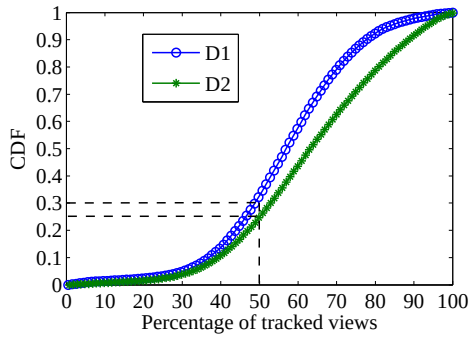


Figure 2: Distribution of percentage of tracked views.

3.2 Main Sources of Overall Views

To figure out the main sources of overall views on YouTube, we investigate the view statistics & data. The percentage of views from each category of view sources is shown in Figure 4, which clearly shows that YouTube Search and Related Video are the top two categories. Views from YouTube Search and Related Video together account for 66.88% and 56.06% of views in D1 and D2, respectively, while views from sources outside YouTube such as Google search, Google video search, Facebook and other sites only account for 7.6% and 5.6% of views in D1 and D2, respectively. From the figure, the percentage of views from Featured in D2 is much higher than that in D1. This is probably due to the fact that D2 was crawled with featured videos as the initial set. Comparing the contribution made by Related Video and YouTube Search, the results show that Related Video contributed a bit less views (2.6% in D1 and 1.2% in D2) than YouTube Search.

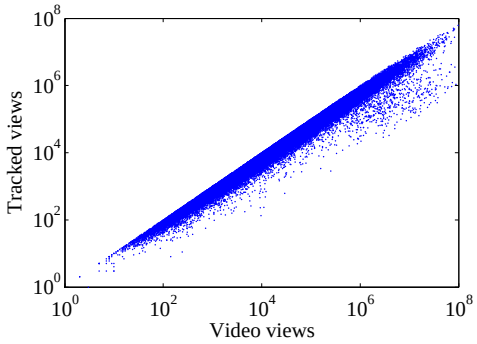


Figure 3: Video views vs. tracked views.

In conclusion, the related video recommendation is one of the main sources of the video views. It accounts for about 30% of the overall views on YouTube, and is only second to the YouTube Search by a small percentage.

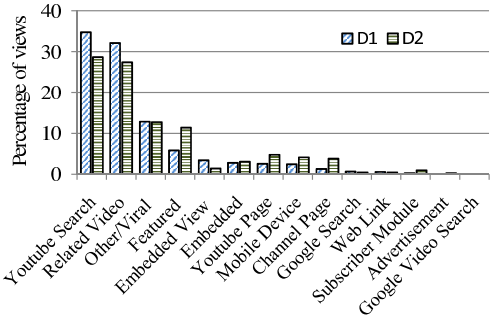


Figure 4: Percentage of tracked views from each category.

3.3 Dominant Sources for Individual Videos

Besides examining the main sources of overall views, we also investigate view sources for each individual video. For each video, we determine the category which contributes the largest proportion of views, and call it the *dominant category*. For example, if the YouTube Search contributes the largest proportion of views to a video among all sources, then YouTube Search is the dominant category of the video, in other words, the video is *dominated* by YouTube Search. The percentage of videos dominated by each category is shown in Figure 5. In contrast to Figure 4, Figure 5 shows

that more videos are dominated by Related Video than YouTube Search. To show this clearly, we aggregate videos into three types, which are Related dominated, Search dominated, and Others. Table 3 shows the percentage of videos of each type for D1 and D2. We can see that the percentage of videos dominated by Related Video is the largest for both data sets.

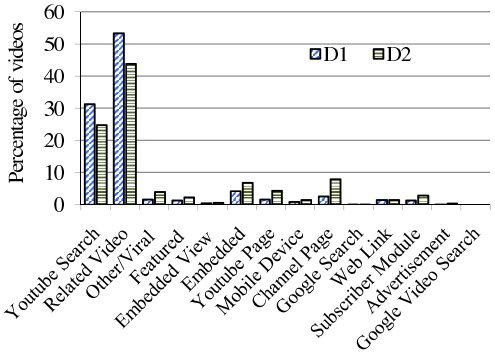


Figure 5: Percentage of videos dominated by each category.

For the disparity between the dominant category of overall views and individual video views, we further investigate the union set (D1&D2) of D1 and D2 to find out whether the dominant categories are different for videos with different popularity. We first aggregate videos into groups with view count range of one thousand views. Figure 6 shows the number of videos of each type for different view count ranges. Among the unpopular videos (left part of the figure), the number of Related dominated videos is the largest among the three types.

Category	D1 (%)	D2 (%)	D1&D2 (%)
Related Video	53.31	43.96	50.83
YouTube Search	31.18	24.71	28.79
Others	15.51	31.33	20.38

Table 3: Percentage of videos dominated by the three types.

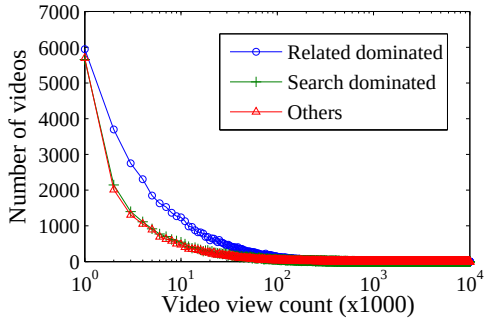


Figure 6: Number of videos for each type after aggregation with constant view count.

To understand which source is the dominant type for popular videos, we aggregate the videos using different view count range. Figure 7 shows the number of videos in each type for different view count ranges. From the figure, most

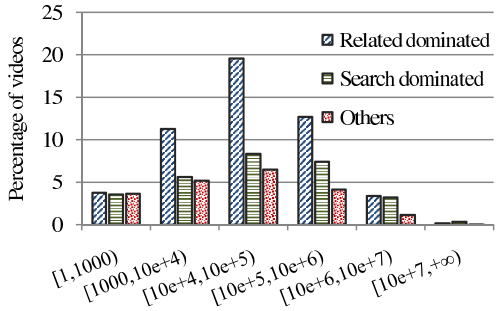


Figure 7: Number of videos for each type after aggregation with different view count.

videos have the view count in the range of one thousand to one million, and the number of Related dominated videos is distinctly the largest in this range. Similarly, we aggregate the videos by their view rates with the constant view rate range of 100 views per day and different view rate range. As shown in Figure 8 and 9, the number of Related dominated videos is always the largest, except for the extremely popular videos which account for 1.2% of total videos.

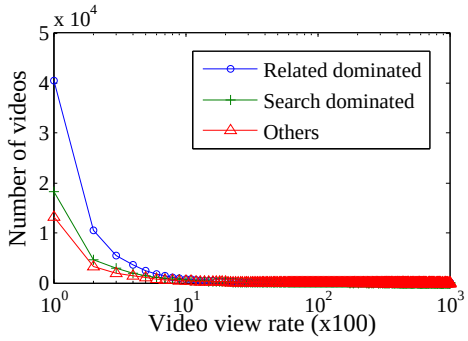


Figure 8: Number of videos for each type after aggregation with constant view rate.

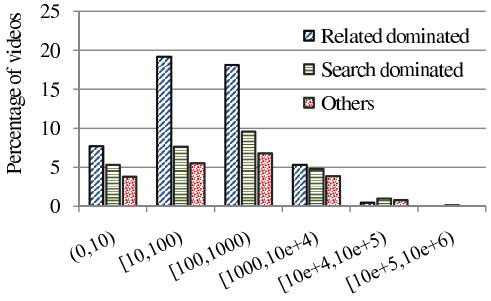


Figure 9: Number of videos for each type after aggregation with different view rate.

Based on the results of both view count and view rate, we conclude that the category of Related Video drives more views than other category of view sources for the majority of videos.

4. HOW RECOMMENDATION SYSTEM AFFECTS VIDEO VIEWS

In this section, we study how the recommendation system on YouTube affects the views of the videos. First, we look at the correlation of views and view rates between a video and its referrer videos. Then, we investigate how the positioning on the related video list affects the view propagation between the videos.

4.1 The Correlation Between Video Views and Referrer Video Views

As shown in Section 3.3, referrer videos are the dominant sources of views for the majority of videos. This suggests that there might be a strong correlation between the view or view rate of a video and its referrer videos. To verify this conjecture, we investigated the correlation of the view count of a video and the average view count of its top referrer videos. The *top referrer videos* of a video is a set of referrer videos that are the major view contributors to the video and are included in the top ten view sources list of the video. We further investigated the correlation of view rate of a video and the average video rate of its top referred videos.

In Figure 10 and Figure 11, we show the correlation between a video and its top referrer videos for view count and view rate, respectively, for the union set of D1 and D2. We observe the trend that the higher the average view count of the top referrer videos, the higher the view count of the video. The same trend can be observed for the view rate, but not as strong as that of the view count. This might be caused by the age difference of a video and its referrer videos.

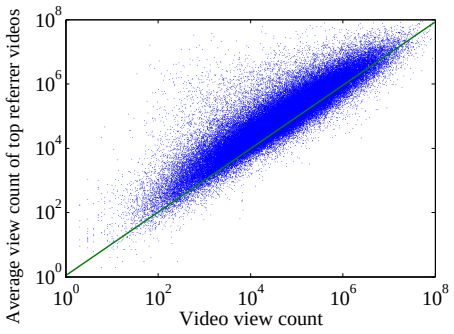


Figure 10: The view count of a video vs. the average view count of its top referrer videos.

We further investigate the correlation by computing the correlation coefficient of views and view rate between a video and its top referrer videos. The correlation coefficients are 0.60 and 0.44 for view count and view rate, respectively. The relatively high correlation coefficients are the additional evidence that the recommendation system has great impact on the views of videos. The implications from these results are that the group of referrer videos of a certain video is a good estimator and indicator for the view count of the video, and a video has a higher chance to become popular when it is placed on the related video lists of popular videos.

4.2 How the Position Affects Video Views

The related videos generated by the recommendation system are placed in order from the top to the bottom of the related video list shown on a video watch page. The understanding of how the position of a video in the related video list affects the click through rate (CTR) is important for understanding how videos drive views for their related videos.

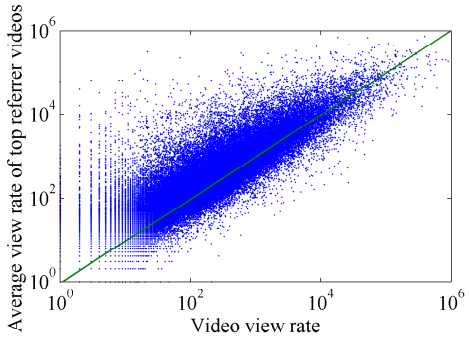


Figure 11: The view rate of a video vs. the average view rate of its top referrer videos.

We use the view statistics & data and the related video lists of referrer videos in D2 to investigate the CTR between related videos. For each pair (V, R) of a video V and its referrer video R, we figure out the position of V on R's related video list, the number of views that R contributed to V , and the number of views that R got after it contributed the first view to V. Then, we calculate the CTR for each pair (V, R), where V is in the P^{th} position of R's related video list ($1 \leq P \leq 20$). Finally, we compute the median of CTRs for each position P.

The result is shown in Figure 12. The figure shows that the CTR declines steadily in a log-log plot as the position lowers, and it fits the Zipf function $Y = \beta X^{-\alpha}$ quite well. The sum of the CTRs from all positions is 41.6%, with the first position having a CTR of 5.9% and the last position 1.0%. This means when a user views a certain video, there is a 41.6% chance that she will watch one of the related videos. The difference between the CTRs of the first and the last position indicates that the position in a related video list plays an important role in video view propagation. However, the CTR for YouTube videos decreases at a slower rate than that of Google advertisement, which was reported that the first position attracted about 39 times more clicks than the 10^{th} position [1]. This reflects the difference in the usage patterns of the recommendation system on the video sharing sites and the web advertisement.

5. IMPACT ON VIDEO VIEW DIVERSITY

In this section, we investigate the impact of YouTube recommendation system on view diversity to understand whether the recommendation system helps users to discover videos of interest but not necessarily popular, or is more likely to recommend popular videos only.

To measure the view diversity, we use the Gini coefficient to measure the distributional inequality [5]. Ordering the videos ascendingly based on their number of views, the Gini coefficient for our case can be computed by

$$G = 1 - \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k + Y_{k-1}),$$

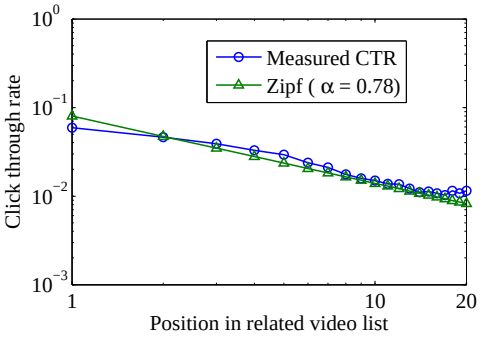


Figure 12: Click through rate for each position in related video lists.

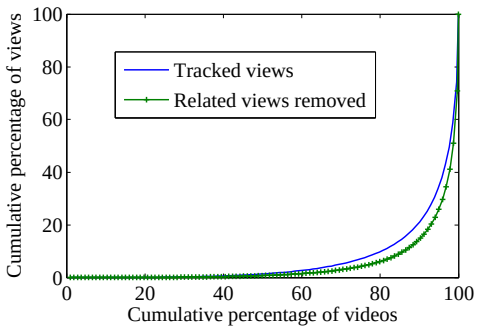


Figure 13: The Lorenz curves of the union set of D1 and D2.

where n is the number of total videos, X_k is the percentage of the first k videos, and Y_k is the percentage of the views from the first k videos.

We compare the Gini coefficients of two scenarios: where the recommendation system is presented and where it is not presented. For the first scenario, we simply use the tracked views to compute the Gini coeficients for data set D1, D2, and the union set of D1 and D2. For the second scenario, where the recommendation system is not presented, we remove the views contributed by referrer videos from the tracked views, and then recalculate the Gini coefficients.

The Gini coefficients and the Lorenz curves for the two scenarios are shown in Table 4 and Figure 13, respectively. From the table, we find that the Gini coefficient for each scenario is quite high, which means the gap between the views of popular videos and niche videos is huge. This is confirmed by the Lorenz curve, which shows more than 90% of the views are from 20% of the videos. Furthermore, when we remove the views from the recommendation system, the Gini coefficient becomes even higher. This indicates that the existence of the recommendation system helps to decrease the Gini coefficient or, in other words, increase the view diversity.

Dataset	Original views	Related removed
D1	0.87	0.90
D2	0.88	0.90
D1&D2	0.87	0.90

Table 4: Gini coefficient.

6. RELATED WORK

As one of the most popular and largest video sharing websites, there are several studies on the YouTube usage pattern and characteristics of videos on YouTube. In [7], Gill et al. study the traffic of YouTube from network edges, providing the insight on YouTube utilization. They analyze the traces to investigate the characteristics of user session on YouTube. Zink et al. also perform the analysis on YouTube usage based on the campus network traces in [12]. The relationship of related videos on YouTube has been studied by Cheng et al in [4]. Their result shows that the related video graph on YouTube exhibits small-world characteristics and has a large clustering coefficient. In [3], Cha et al. study various aspects about video popularity and video view on YouTube. They propose to use the history of video view to predict video popularity in near future. Our work compliments their work by showing that there is a correlation between a video's view count and its related videos' view count, and this should be another factor to consider in predicting video popularity.

Several previous works study the impact of recommendation system on the user behavior and the improvement for the recommendation system. In [5] and [6], Fleder and Hosanagar use analytical model and simulation to study the impact of the recommender on sales diversity. In [11], Zhou et al. propose the recommendation systems that aim to achieve both diversity and accuracy.

The propagation of video views to its related videos is related to influence spreading in social networks. Leskovec et al. study the influence patterns in user-to-user recommendation network in [8]. Zhao et al. present several mathematical models of influence spreading in [10].

7. CONCLUSION

In this paper, we perform a measurement study on the impact of related video recommendation system on video views. Through the measurement of view sources, we find that the related video recommendation accounts for about 30% of overall views. Moreover, it is the most important view source for the majority of videos. By investigating how video views are driven by the recommendation system, we find a strong correlation between the view count of a video and the average view count of its top referrer videos, and also discover that the position of a video on a related video list plays a critical role in the click through rate of the video. Finally, the evaluation of the impact of the video recommendation system on the diversity of video views shows that the existence of YouTube recommendation helps to increase the diversity of video views in aggregation, which means that YouTube recommendation helps viewers discover more videos of their interest rather than the popular videos only.

8. ACKNOWLEGEMENTS

The authors are grateful to the anonymous reviewers for their helpful comments and suggestions. This work is partially supported by U.S. NSF grants CNS-066618. Renjie Zhou was a visiting student at University of Massachusetts, Amherst, supported by China Scholarship Council, when this work was performed.

9. REFERENCES

[1] Google adwords click through rates per position. `http://www.accuracast.com/seo-weekly/adwords-clickthrough.php`, October 2009.

[2] S. Axon. Youtube to let users charge rental fees. `http://www.cnn.com/2010/TECH/05/03/youtube.rental/index.html`, May 2010.

[3] 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 *Proceedings of ACM Internet measurement Conference(IMC), San Diego, CA, USA*, October 2007.

[4] X. Cheng, C. Dale, and J. Liu. Statistics and social network of youtube videos. In *International Workshop on Quality of Service (IWQoS'08)*, pages 229–238. IEEE, June 2008.

[5] D. Fleder and K. Hosanagar. Recommender systems and their impact on sales diversity. In *Proceedings of the 8th ACM conference on Electronic commerce*, page 199. ACM, 2007.

[6] D. Fleder and K. Hosanagar. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Science*, 55(5):697–712, 2009.

[7] P. Gill, Z. Li, M. Arlitt, and A. Mahanti. Characterizing Users Sessions on YouTube. In *Proceedings of SPIE/ACM Conference on Multimedia Computing and Networking (MMCN), Santa Clara, USA*, January 2008.

[8] J. Leskovec, A. Singh, and J. Kleinberg. Patterns of influence in a recommendation network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD*, pages 380–389. Springer-Verlag, 2005.

[9] P. White. How many videos are on youtube? `http://www.associatedcontent.com/article/1927414/how_many_videos_are_on_youtube.html?cat=15`, July 2009.

[10] B. Zhao, Y. Li, J. Lui, and D. Chiu. Mathematical Modeling of Advertisement and Influence Spread in Social Networks. In *NetEcon:Workshop on the Economics of Networks, Systems and Computation*, 2009.

[11] T. Zhou, Z. Kuscsik, J. Liu, M. Medo, J. Wakeling, and Y. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511, 2010.

[12] M. Zink, K. Suh, Y. Gu, and J. Kurose. Watch Global, Cache Local: YouTube Network Traffic at a Campus Network - Measurements and Implications. In *Proceedings of SPIE/ACM Conference on Multimedia Computing and Networking (MMCN), Santa Clara, USA*, January 2008.