A Study of YouTube recommendation graph based on measurements and stochastic tools[*,*]

Yonathan Portilla[†,†], Alexandre Reiffers[†] [‡,‡], Eitan Altman[⋆], Rachid El-Azouzi[†]

ABSTRACT

The Youtube recommendation is one the most important view source of a video [1]. In this paper, we focus on the recommendation system in boosting the popularity of videos. We first construct a graph that captures the recommendation system in Youtube and study empirically the relationship between the number of views of a video and the average number of views of the videos in its recommendation list. We then consider a user that browses through videos such that the video it chooses to watch a given time is selected randomly among the videos in its recommendation list. We study the stability properties of this random process and we show that the trajectory obtained does not contain cycles if the number of videos in the recommendation list is small (which is the case if the computer's screen is small).

1. INTRODUCTION

Online media constitute currently the largest share of Internet traffic. A large part of such traffic is generated by platforms that deliver user-generated content (UGC). This includes, among the other ones, YouTube and Vimeo for videos, Flickr and Instagram for images and all social networking platforms. Recently, Youtube becomes more popular and none have achieved the same success. Based on statistics available from the website Alexa.com, more than 30% of global internet users visit Youtube.com per day. Other statistics from:

http : //www.youtube.com/t/press_statistics clearly illustrate the previous fact: "Over 800 million unique users visit YouTube each month" and "72 hours of video are uploaded to YouTube every minute". Of course, not all videos posted on YouTube are equal. The key aspect is their "popularity", broadly defined as the number of views they score (also referred to as view count). This is relevant from a twofold perspective. On the one hand, more popular content generates more traffic, so understanding popularity has a direct impact on caching and replication strategy that the provider should adopt. On the other one, popularity has a direct economic impact. Indeed, popularity or view count are often directly related to click-through rates of linked advertisements, which constitute the basis of the YouTube's business model. Hence the revenue model of YouTube is based on a sophisticated advertising scheme. Indeed different types of advertising methods are used in YouTube, for example, "In-video graphical and text advertisements", "post-roll advertising", etc. Rising incomes in such business models are related to content visibility, which motivates YouTube to provides the recommendation list and to display features videos.

Models for predicting popularity of online content including YouTube videos and Digg stories, are proposed in [2, 3, 4, 5, 6, 7], with the aim of developing models for early-stage prediction of future popularity [8]. Such studies have highlighted a number of phenomena that are typical of UGC delivery. This includes the fact that a significant share of content gets basically no views [7], as well as the fact that popularity may see some bursts, when content "goes viral" [5]. Visibility of content is not just of interest to Youtube. It is also of interest to the content creators. There is a competition between them on visibility of their creations. Understanding of how view count of a video is driven by different sources of views is helpful for finding strategies for increasing the number of views of videos. For advertisers and for content providers, this is useful for strategic planning so as to increase the contents' popularity. This is often directly related to click-through rates of linked advertisement.

To achieve this, we will study properties of recommendation lists and their impact on content propagations. Several studies have showed that there exists a strong correlation of view count of a content and the average view count of its recommendation list [1, 9]. Indeed, from different measurement studies, the view count of videos in a recommendation list of a video tends to match the view count of that video. Cheng et al. have provided different statistics of Youtube and showed that the graph of YouTube'video structure exhibits a small-word characteristic [9]. This observation is expected, due to the user-generated nature of the tags, ti-

[*]This work has been supported by the European Commission within the framework of the CONGAS project FP7-ICT-2011-8-317672, see www.congas-project.eu.

[†⋆]INRIA B.P.93, 2004 Route des Lucioles, 06902 Sophia-Antipolis, Cedex, FRANCE

[‡†]CERI/LIA, University of Avignon, 74 rue Louis Pasteur, 84029 AVIGNON Cedex 1

tle and description of the videos that are used by YouTube to find related videos to appear in the recommendation list. Zhou et al. have showed the importance of the recommendation system on view count of a video [1]. They found that the recommendation system is the source for 40-60% views of a video. Performing several measurement, they have discovered that the position of a video on a related video plays an important role in the click through rate of the video. They have also identified that the recommendation system improves the diversity of video views which helps a user to discover videos with less popularity.

Our work compliments their work by quantifying the relationship between a video's view and its related videos (i.e. the videos in its recommendation list). To that end we focus on users that browse through videos according to some random mobility model over the recommendation graph. In this directed graph, videos are nodes, and directed edges connect that node to the videos that appear in its recommendation list. Nodes have some attributes, or weights.

Our goal here is to relate the attributes of a node to those of its recommendation graph, We then study properties of the sequence of weights in a random trajectory of a user in the recommendation model as a function of their mobility model. In particular, we focus on two attributes. The first is the number view count of the video, and the second is its age. To each one of these attributes and every mobility model, the sequence of consecutive attributes on the random trajectory forms a stochastic process which we model as a Markov chain and we study its stability. We then derive properties related to the stability of the sequence of videos viewed from the stability properties of the Markov chain corresponding to the attribute process. Finally, we also provide a theoretical result on the improvement by the recommendation system of the diversity of video views. Indeed, recommender system in Youtube is traditionally based on keyword between videos (tags, title and summary). In order to increase the diversity and suggest best long tail videos to the user, the system needs to provide recommendation depends not only on textual description. In this study we will identify how the level of the popularity of a content has an impact on the related video selected by Youtube recommendation system and how this correlation can have an impact on video view diversity.

The remainder of the paper is structured as follows. The recommendation graph of you tube is introduced in the next section, along with a statistic analysis of the relation between the number of views of a video and the number of views of those in its recommendation graph. We study the stability of this random process in section 4. Section 5 concludes the paper.

2. A MODEL FOR YOUTUBE RECOMMENDATION SYSTEM

We now provide some background onYoutube which becomes a key international platform for socially enabled media diffusion. This platform allows not only to share videos, but also to create interaction between users (friends, creators, rating..). It becomes a most attractive and a popular media diffusion with a huge quantity user-generated content. Our study on recommendation system in Youtube is based on the data sets crawled from Youtube. Here we describe how we collected the data sets.

2.1 Data on videos

All youtube videos are available to the general public and includes valuables data about the video. Moreover, Youtube proposes a list of recommendation which contains the related videos that the system recommend for a user watching a video from youtube platform. In this subsection, we describe how we collect data sets using our software developed in JAVA. This tool allows us to collect some view statistic of videos in Youtube as views, titles, tags, ages and recommendation list. A Linear Least Squares Regression (LLSR) is used to adjust the model parameter in order to obtain the minimum error between the model and experimental data.

2.2 View graph

In this section we construct a graph based on Youtube recommendation. In particular, we will explore how the view of a video influences the view count of other videos through the recommendation list. Indeed, a user who views a video u may view a video v from its recommendation list. In that case, there is a directed edge between u and v (See figure 1).

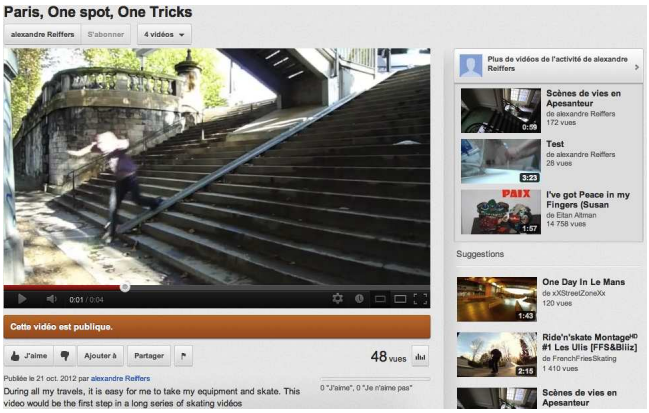


Figure 1: Recommendation list in Youtube

Let us first introduce some terminology. We consider a connected network $G = (V, E)$, where V denotes the set of nodes with $|V| = n$, E denotes the set of edges and $w : V \to \mathbb{R}_+$ denotes the view count of the node v. We will now describe how we construct our connected graph. We imagine a random walker on Youtube starting from a video u. After viewing the video u, he selects randomly (with uniform distribution) a video v among top N videos in its recommendation list, and moves to its neighbor. We repeat the procedure again and again. A stochastic (transition) matrix $Q = [Q_{uv}]_{n\times n}$ is used to govern the transition of the random walk process where n is the number of videos in the graph. Q_{uv} is the probability that the transition from video u to video v occurs.

3. STATISTICAL STUDY OF THE RECOMMENDATION GRAPHS

In this section, we explain how we obtain the influence of recommendation system using our data sets in Youtube. We

Table 1	regression coefficients	R^2
$N = 1$ using logarithmic scale	1.03932	0.746919
$N = 1$ using linear scale	0.658874	0.025083
$N = 2$ using logarithmic scale	0.986524	0.987252
$N = 2$ using linear scale	2.73843	0.773921
$N = 3$ using logarithmic scale	1.00942	0.825262
$N = 3$ using linear scale	0.114771	0.005682
$N = 4$ using logarithmic scale	1.01938	0.816736
$N = 4$ using linear scale	0.858309	0.034767
$N = 5$ using logarithmic scale	0.905340	0.791218
$N = 5$ using linear scale	2.73843	0.221205
$N = 8$ using logarithmic scale	0.587379	0.335723
$N = 8$ using linear scale	0.798919	0.176054
$N = 10$ using logarithmic scale	0.529770	0.277402
$N = 10$ using linear scale	2.99831	0.195627
$N = 15$ using logarithmic scale	0.500745	0.258681
$N = 15$ using linear scale	1.21020	0.069401

Table 1: Regression coefficients and coefficient of determination R-square for different values of N

randomly picked 1000 videos and we focus on two elements of the data. The first element is the view count of a video and the second element is the average view count of related videos recommended by YouTube recommendation system. We believe that the number of videos that we collected for each experiment is enough to capture all information. First, we investigate the relation between the view count of a video and the average view count of its recommendation list. We consider different values of the number N of videos in the list. In practice, the larger screen is, the larger is N. We shall show later that N plays a crucial role on the properties of the excursions over the recommendation graph. A very small N corresponds to list of recommendations viewed over cellular telephones.

In our statistical study, we test two types of model that relate the number x_i of views of a video i and the number y_i of views of videos in its recommendation list averaged over N:

- N-videos in which we use the linear regression between x_i and y_i, $i \in \{1, \ldots, 1000\}$.
- N-videos wherein a linear regression is used between $log(x_i)$ and $log(y_i)$.

In figure 3, we plot the view count of one video on the X-axis, and the average view count of videos in its recommendation list. We observe that the coefficient of determination R-square is small for all values of N (see Table 3) . This indicates that the regression line cannot explain all the variation of the average of view count of videos in recommendation list of a video by variation in view count of that video. In other hand, in figures 3-3, we use a logarithmic scale to plot the view count of one video on the X-axis, and the average view

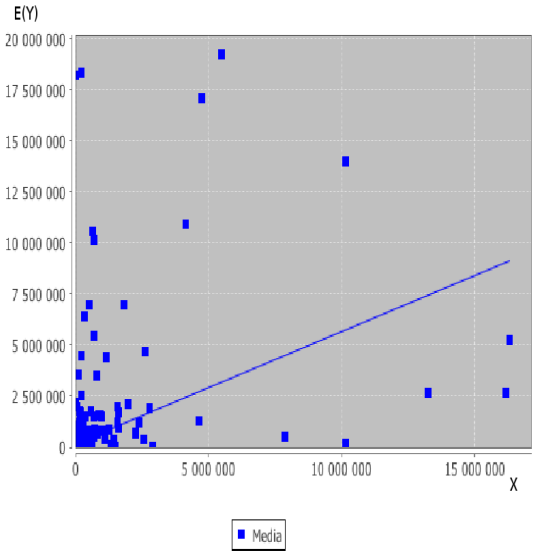


fig1

(a) N=3

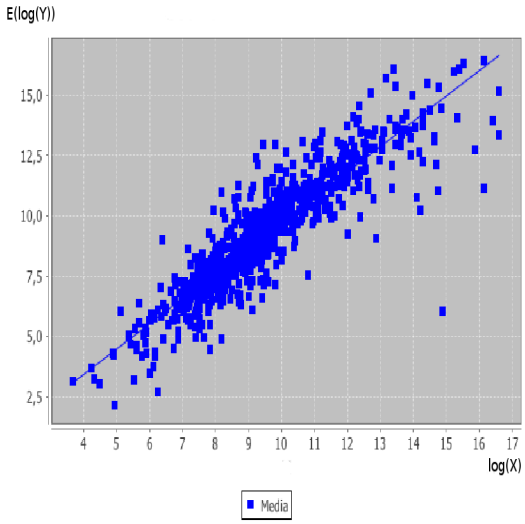


fig2

(b) N=3

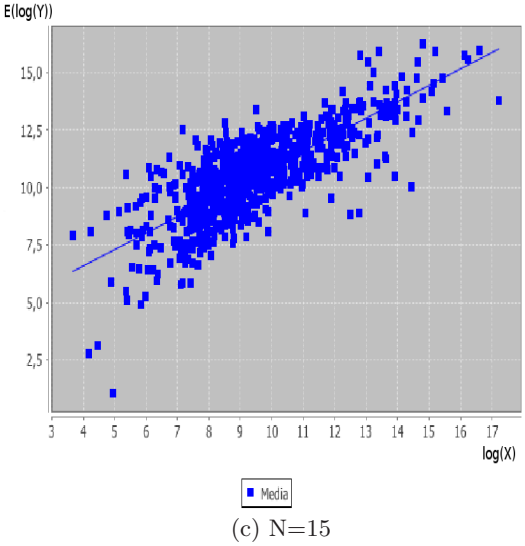


fig3

(c) N=15

Figure 2: The view count of one video on the X-axis, and the average views of the top $N = 3,\ 15$ of its recommendation list

count of videos in its recommendation list. Both figures 3-3 show the strong correlation between the view count of a video and its average of view count of top N videos in its recommendation system. The first observation claims that the higher average of view count of related videos in recommendation list of a video, the higher view count of that video. Hence, Youtube recommendation system will prefer to put videos in recommendation list based on the popularity of the current video, located in the same region of the popularity (same region of the 'long tail' of popularity). On the other hand, we characterize a good relationship between the view count x of a video and the average of view count y of videos in its recommendation list. This relationship is $y = (e^{\ln(x)})^a + e^b$. The table 1 shows how well a regression line using logarithm scale, fits a set of data for several values of N. This approximation is more accurate when N is small. Since a walker has more probability to choose a video at top position (Google advertisement [10] showed that the first position in the recommendation list attract 39 times more click that the 10^{th} position), this relationship is still a very good approximation even if the number of videos in list of recommendation system is high.

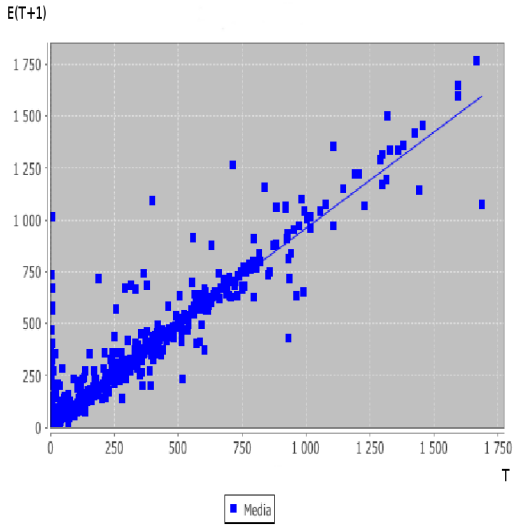
Now we further investigate the correlation between the age of a content and the average of age of videos in its recommendation list. Figures 3-3 show clearly the trend that the higher the average of age of the related videos, the higher the age of the video. This implies that Youtube recommendation will prefer to recommend the videos located in the same region of the age. This means that Youtube's recommendation will not significantly affect the overall videos based on the age of a video and will instead focus more on the same generation of that video. Moreover, the table 3, provide a high regression coefficient which is the additional evidence that the recommendation system has an important impact on the view count and a popular video can affect only the videos located in the same region of the popularity and the age.

4. STABILITY AND VIDEO VIEW DIVERSITY

Consider X_n, $n \geq 1$ a sequence of random process that satisfies the Markov property and takes values from a set $S = \{1, 2, .., n\}$ where n is the view count of the visited video. We construct this random process by using a random walker moving in the graph as defined in the previous section, but we take into account only the view count of the video.

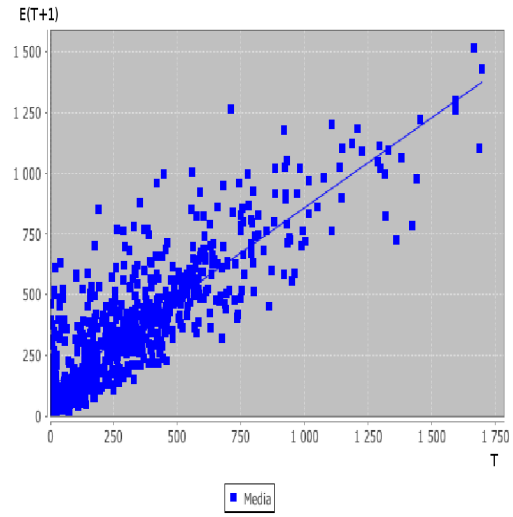
In this section, we investigate the stability of Markov chain defined above in order to identify the impact of Youtube recommendation system on view diversity and how suggest best long tail videos that are undiscovered because of its less view count. We are interested to the stability in order to show if the random walker can remain in a small region starting from an initial video and move to other video using YouTube recommendation system. In order to study the stability we will define the first hitting time as follows

Definition 1. *We define the hitting time of a set of states*



(a) N=3

fig4



(b) N=10

fig5

Figure 3: The age of one video on the X-axis, and the average age of the top $N = 3, 15$ of its recommendation list

Table 3	regression coefficients	R2
1 videos days	0.976898	0.918683
2 videos days	0.932103	0.882182
3 videos days	0.931216	0.879587
4 videos days	0.906025	0.870077
5 videos days	0.906025	0.870077
8 videos days	0.780917	0.776079
10 videos days	0.742248	0.713315
15 videos days	0.723398	0.685777

(a)

table2

Figure 4: Regression coefficients and coefficient of determination R-square for different values of N

B by

$$\tau_B = \inf\{n \geq 1 : \ X_n \in B\}$$

This set is called a positive recurrent if

$$\sup_{x \in B} E_x \tau_B < \infty$$

The Markov chain is stable if it is a positive class recurrent.

From different experiments (see for example Table 3), we observe that for $N = \{3, 4, 5\}$ and for all $i \in \mathbb{N}^*$ we have

$$\mathrm{E}\big[\log(X_{n+1}) \mid X_n = x\big] < \infty, \quad \text{and}$$

$$\mathrm{E}\big[\log(\mathrm{X}_{n+1}) - \log(X_n) \mid X_n = x\big] \geq 0 \text{ for } \log(x) \geq K > 0$$

It follows from Theorem 3 in [11], that he Markov chain cannot be positive recurrent.

The instability of the Markov chain for $N = \{3, 4, 5\}$ (which is the case if the computer's screen is small) has several interpretations and consequences. Indeed, by the rapid adoption of smartphone, tablets and e-reader which are characterized by a small screen, our result may explain some results in [1] which showed how Youtube recommendation can improve the view diversity and help users to discover videos in long tail.

5. CONCLUSION

This paper has several contributions. Firstly we showed through the measurements the relation between function of views of a video and the function averaged over the number of views of videos in its recommendation list. Based on this relationship we explored the evolution on number of views that a random walker sees when traveling through recommendation graph. We showed in particular if the number of videos in the list is small, the number of views tends to increase along the trajectory of random walker. We conclude that the random trajectory defined as Markov chain, is not sable which means that the trajectory does't not contain cycles.

6. REFERENCES

[1] R. Zhou, S. Khemmarat, and L. Gao, "The impact of youtube recommendation system on video views," in *Proc. of IMC 2010*, Melbourne, November 1-3 2010.

[2] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proc. of ACM IMC*, San Diego, California, USA, October 24-26 2007, pp. 1–14.

[3] R. Crane and D. Sornette, "Viral, quality, and junk videos on YouTube: Separating content from noise in an information-rich environment," in *Proc. of AAAI symposium on Social Information Processing*, Menlo Park, California, CA, March 26-28 2008.

[4] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube traffic characterization: A view from the edge," in *Proc. of ACM IMC*, 2007.

[5] J. Ratkiewicz, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani, "Traffic in Social Media II: Modeling Bursty popularity," in *Proc. of IEEE SocialCom*, Minneapolis, August 20-22 2010.

[6] G. Chatzopoulou, C. Sheng, and M. Faloutsos, "A First Step Towards Understanding Popularity in YouTube," in *Proc. of IEEE INFOCOM*, San Diego, March 15-19 2010, pp. 1 –6.

[7] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, pp. 1357 – 1370, 2009.

[8] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Comm. of the ACM*, vol. 53, no. 8, pp. 80–88, Aug. 2010.

[9] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of youtube videos," in *International Workshop on Quality of Service*, June 2008.

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

[11] S. Foss, "Coupling again: the renovation theory," *Lectures on Stochastic Stability, Lecture 6, Heriot-Watt University*. [Online]. Available: http://web.abo.fi/fak/mnf/mate/tammerfors08/Foss_Lecture6.p