A Study of YouTube recommendation graph based on measurements and stochastic tools

Yonathan Portilla, Alexandre Reiffers, Eitan Altman, Rachid El-Azouzi

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. 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 random walker on the recommendation graph, i.e. a random user that browses through videos such that the video it chooses to watch is selected randomly among the videos in the recommendation list of the previous video it watched. 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).**

***Index Terms*— Analysis of recommendation system, drift stability analysis, Youtube**

I. 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, YouTube and Vimeo for videos, Flickr and Instagram for images and all social networking platforms. Among these platforms, Youtube has become the most popular. 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 hand, popularity has a direct economic impact. Indeed, popularity or view counts 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. Extracting incomes in such business models is related to content visibility, which motivates YouTube to provide 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], [8], with the aim of developing models for early-stage prediction of popularity features [9]. 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 consumers and to the content creators. There is a competition between the creators 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 propagation. Several studies have showed that there exists a strong correlation of view count of a content and the average view count of the videos in its recommendation list [1], [10]. Indeed, from different measurement studies, the view count of videos in a recommendation list of a given 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's video structure exhibits a small-world characteristic [10].

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 unpopular videos.

Our work compliments these works 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

*INRIA B.P.93, 2004 Route des Lucioles, 06902 Sophia-Antipolis, Cedex, FRANCE

† C.E.R.I., University of Avignon, France

that may correspond to the view count, or to the total time that the video was viewed, or to the number of likes etc.

We first describe the model associated the recommendation system in section II. Our first goal is to relate the attributes of a node to those of its recommendation graph. In section III, we first describe our dataset. We then study empirical 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 in section IV. 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 in section V, we 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).

II. A MODEL FOR YOUTUBE RECOMMENDATION SYSTEM

We now provide some background on Youtube which has become 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 of 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.

A. Data on videos

A huge majority of Youtube videos are available to the general public and includes valuables data about the video (a video may not be available if its creator decided to attach to it the unlisted or the private privacy level or if there are copyright issues that do not allow to show the video at a given country). Youtube further proposes a list of recommendation from which the consumer can choose the next video to watch. We collected data sets of videos using our own 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.

B. 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).

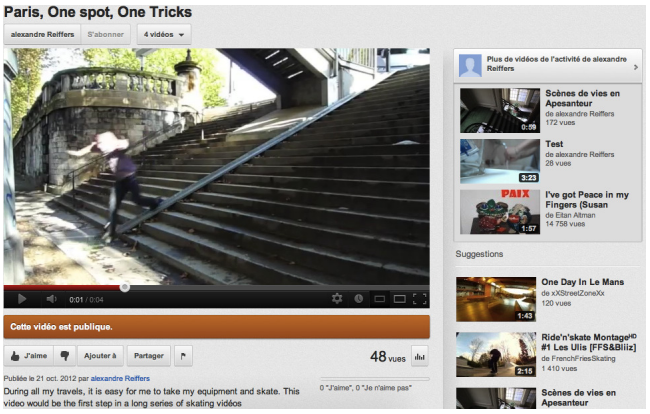


Fig. 1: Recommendation list in Youtube

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 neighbour. 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.

III. 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 will study two datasets. In the first one we randomly picked 1000 videos, in the month of july 2012 in Youtube by using our random extraction software. In the second one we crawl Youtube to get *NBVIDEOS* videos. 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 models that relate a function of the number x_i of views of a video i and

the function 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, $i \in \{1,\ldots,I\}$, wher $I \in \{1000, NBVIDEOS\}$. Here,

$$y_i = \frac{1}{N}\sum_{j=1}^{N} y_i^j$$

where y_i^j is the view count of the j^{th} video in the recommendation list of video i. We thus identify parameters a and b so that y_i will be well approximated by $ax_i + b$. a and b are chosen in fact so as to minimize the mean square error of this approximation over all samples.

- N-videos wherein a linear regression is used between $log(x_i)$ and $log(y_i)$. We thus identify parameters a and b so that

$$\frac{1}{N}\sum_{j=1}^{N} \log(y_i^j)$$

will be well approximated by $a\log(x_i) + b$. a and b are chosen in fact so as to minimize the mean square error of the difference between the expressions over all samples.

A. *Study of the data set*

The distribution of the dataset according to age and to popularity

The measure of goodness of the linear fit is summarized by the statistical R^2 coefficient. It takes value in the unit interval $[0,1]$ and the better the fit is, the larger is its value.

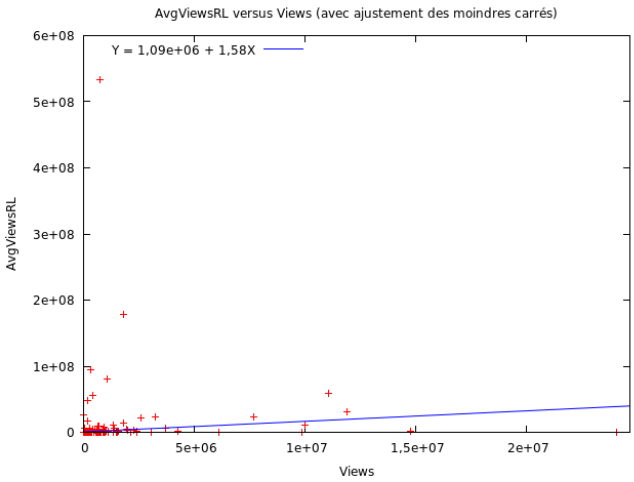
In figure 2a, we plot the view count of one video on the X-axis, and the average view count of videos in its recommendation list. We observe that the coefficient of determination R-square is small for all values of N (see fig. 4a) . This indicates that there are important deviations from any approximation of the y_i as a linear function of the x_i.

On the other hand, in figures 2b-2c, we use a logarithmic scale to plot the view count of one video on the X-axis, and the averaged of the view count of videos in its recommendation list. Both figures 2b-2c show the strong correlation between the view count of a video and the view count of top N videos in its recommendation system.

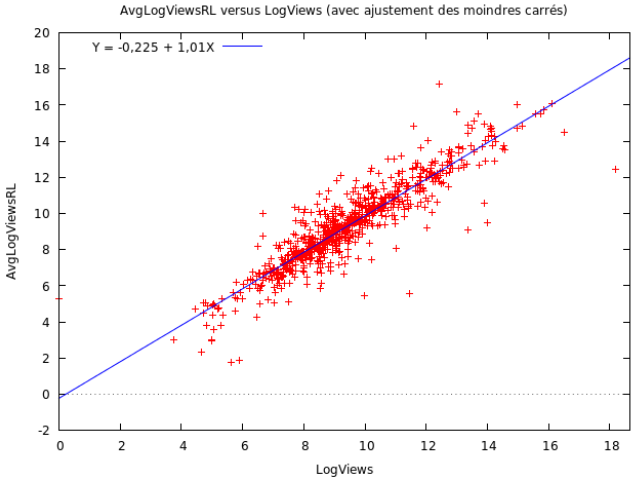
The first observation is on the monotonicity: the higher the the view count of the related videos in recommendation list of a given video is, the higher the view count of that video is. Hence, Youtube recommendation system will prefer to put videos in recommendation list based on the popularity of the current video, located in the same region of the popularity.

Table 1 shows the quality of the linear regression in the logarithm scaling. The table fits a set of data for several values of N. Our experimentation thus correspond to random walkers that after watching a video, it chooses the next one to see among its N recommended ones with equal probability. This can model space limitation on the screen

(a) N=1



(b) N=3



(c) N=5

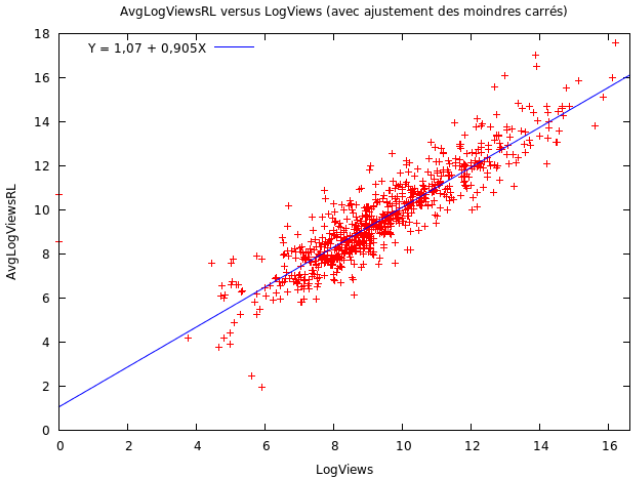
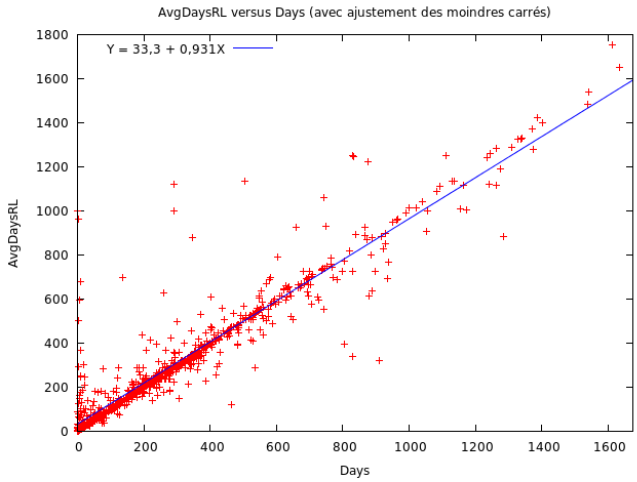


Fig. 2: The view count of one video on the X-axis, and the average of the logarithmic views of the top $N = 1,2,3$ of its recommendation list

(a) N=3



(b) N=10

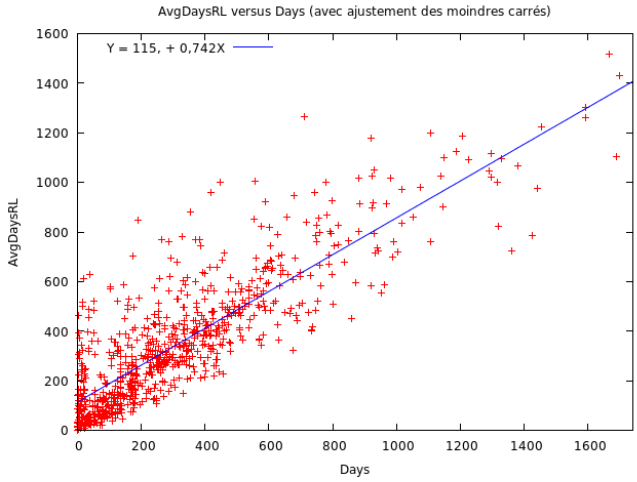


Fig. 3: The age of one video on the X-axis, and the average age of the top $N = 3$, 15 of its recommendation list

or in the recommendation list which limit the number of recommendations seen by the random walker. It can also model user behaviour in which the top recommendations are preferred. Indeed, in was shown in [11] the first position in the recommendation list attracts 39 times more clicks than the 10^{th} position.

Now we further investigate the correlation between the age of a content and the average of age of videos in its recommendation list. Figures 3a-3b 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, Table 4b displays a high regression coefficient which is the additional evidence that the recommendation system has an important

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 2	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

Fig. 4: Regression coefficients and coefficient of determination R-square for different values of N for the age study

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.

IV. Stability and Video View Diversity

We shall study in this section the stochastic process X_n, $n \geq 1$ of the attributes of the videos encountered by the random walker. For example, X_n can be taken to be the view count or the age of the nth visited video. For simplicity we shall assume that this constitutes a discrete time aperiodic communicating Markov chain (any state can be reached from any other state with positive probability), although this assumption will be relaxed later. We investigate the stability of this Markov chain defined above in order to identify the impact of Youtube recommendation system on the visibility and popularity of videos.

We briefly recall some basic notions in Markov chains. The chain is said to be recurrent if it visits every state infinitely often. It is said to be positive recurrent if for each state x, the expected time between consecutive visits of x is finite. It is said to be transient if for some state x there is a strictly positive probability never to return back to that state. Thus in a transient Markov chain, a state x is only visited finitely often, and after some finite (random) time it

never returns to x again. We shall say that X_n is stable if it is positive recurrent and that it is unstable if it is not.

We argue that a recommendation system tought to be designed with the objective of having X_n communicating and positive recurrent for various reasons. First, it would allow a random walker to return to a video he liked. Secondly, it does not get stuck in a subset of states due to the communicating assumption. The fact that one reaches a state infinitely often guarantees in particular that any video will be discovered by any random walker that stays long enough. Finally, this property has a positive impact on increasing the page rank of the video and thus search engines based on page rank will allow one to find the video quicker.

We next recall the Foster stability conditions for communicating Markov chains. Introduce a positive and increasing function f and a finite set B. Define the drift of the chain at state x as

$$\Delta(x) := E[f(X_{n+1}) - f(X_n) \mid X_n = x]$$

Note that due to the Markov property this does not depend on n. If is called a Lyapunov function.

Lemma 1: (i) If $\sup_x \Delta(x)$ is finite and for some $\varepsilon > 0$, $\Delta(x) < -\varepsilon$ for all $x \notin B$ then the Markov chain is stable.
(ii) If $\Delta(x) > \varepsilon$ for all $x \notin B$ then the Markov chain is unstable. A sufficient condition for instability is that

$$E[f(X_{n+1}) \mid X_n = x] \geq af(x) + b$$

for $b > 0$ and $a \geq 1$.

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

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

$$E[\log(X_{n+1}) - \log(X_n) \mid X_n = x] > \varepsilon > 0 \qquad (1)$$

for some $\varepsilon > 0$ and for all x large enough (larger than some constant K). It follows from Theorem 3 in [12], that the Markov chain cannot be positive recurrent.

V. Discussion

The instability of the Markov chain X_n 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.

The process X_n which we studied in this paper can be interpreted as a random walk on the recommendation graph, where X_n corresponds to the number of views of the nth viewed video. The $n+1$ video viewed is chosen uniformly from the recommended videos of the previously viewed video. We identified in this paper instability of the process X_n as a function of N (size of recommendation graph). The instability was obtained by establishing (through experiments) that the related drift is positive.

How sensitive are the results to the Markovian assumption on X_n? Note (by taking expectations in (1)) that as $n \to \infty$, $E[f(X_n)]$ tends to infinity. This type of instability result does not require X_n to be a Markov chain. Some of the stability results of X_n do hold even when X_n is not a Markov chain [13].

Note that both for stability as well as for instability, the drift condition is required to hold for all large enough states, i.e. for all $x > K$ for some positive K. Instability of the process means that the number of views of the nth video tends to grow without bound. In practice however, there is some finite bound on the number of viewed videos (the bound is the number of views of the most viewed video). Thus, what instability means *in practice* is that the number of views grows until it is in some neighbourhood of that bound. For this to occur, we expect that it is sufficient to establish that the drift is bounded below by a positive number within some large integral $x \in [K_1, K_2]$.

VI. Conclusion

We have shown in this paper through measurements the relation between the number of views of a video and the number of views of the videos in its recommendation list averaged over the first N slots in this recommendation list. We considered not only relations between the numbers of videos viewed but also between functions of these numbers. More precisely, we established the relation between a function of the number of views of a video and the function averaged over the number of views of videos in the recommendation list of the video.

We show good fits of a linear relation between the number of views of a video and the number of views averaged over its recommended videos when considering the log-log scale. The fit was not good when considering a linear scale.

Based on this relationship we explored the evolution on number of views that a random walker sees when travelling through recommendation graph. We showed in particular that 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 is not sable which means that the trajectory doest not contain cycles.

For showing instability, we used Foster Criteria where we approximated conditional expectations by conditional averages. Note that Foster Criteria is valid independently of the value of the R^2 parameter.

This study was restricted to Youtube and based on a dataset that reflects Youtube's protocol at a given time (july 2012). However the methodology is valid for other recommendation systems as well or for newer versions of Youtube recommendation.

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] C. Richier, E. Altman, R. El-Azouzi, T. Jimenez, G. Linares, and Y. Portilla, "Bio-inspired models for characterizing youtube viewcount," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 20)14)*, Beijing, China, Aug 2014, pp. 297–305.

[9] 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.

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

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

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

[13] A. A. Borovkov and V. Yurinsky, *Ergodicity and stability of stochastic processes*. J. Wiley, 1998.