HAL
open science

Bio-Inspired Models for Characterizing YouTube Viewcount

Cedric Richier, Eitan Altman, Rachid El-Azouzi, Tania Jimenez, Georges Linarès, Yonathan Portilla

To cite this version:

Cedric Richier, Eitan Altman, Rachid El-Azouzi, Tania Jimenez, Georges Linarès, et al.. Bio-Inspired Models for Characterizing YouTube Viewcount. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Aug 2014, Beijing, China. pp.297-305, 10.1109/ASONAM.2014.6921600 . hal-01011069

HAL Id: hal-01011069

https://inria.hal.science/hal-01011069v1

Submitted on 22 Jun 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bio-Inspired Models for Characterizing YouTube Viewcount

Cédric Richier*, Eitan Altman†, Rachid Elazouzi*, Tania Jimenez*, Georges Linares* and Yonathan Portilla†
*University of Avignon, 84000 Avignon, FRANCE
Email: firstname.lastname@univ-avignon.fr
†INRIA, B.P 93, 06902 Sophia Antipollis Cedex, FRANCE
Email: eitan.altman@inria.fr

***Abstract*—The goal of this paper is to study the behaviour of viewcount in YouTube. We first propose several bio-inspired models for the evolution of the viewcount of YouTube videos. We show, using a large set of empirical data, that the viewcount for 90% of videos in YouTube can indeed be associated to at least one of these models, with a Mean Error which does not exceed 5%. We derive automatic ways of classifying the viewcount curve into one of these models and of extracting the most suitable parameters of the model. We study empirically the impact of videos' popularity and category on the evolution of its viewcount. We finally use the above classification along with the automatic parameters extraction in order to predict the evolution of videos' viewcount.**

***Keywords*-Online videos, bio-inspired models, video popularity, regression model, popularity growth, popularity prediction.**

I. Introduction

YouTube has been one of the most successful user-generated video sharing sites since its establishment in early 2005. It constitutes currently the largest share of Internet traffic. The rate of subscription to YouTube as well as the rate of submitted videos has been growing steadily ranking YouTube and none of its competitors has achieved a similar success [1], [2]. An important aspect of a videos in YouTube is its popularity, which is defined as the number of views (referred as viewcount). Understanding and predicting the popularity is useful from a twofold perspective: on 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; and on the other hand, popularity has a direct economic impact. A number of researchers have analyzed the popularity characteristics of user-generated video content for understanding the processes governing their popularity dynamics [3], [4], [5], [6], [7], [8], with the aim of developing models for early-stage prediction of future popularity [9]. There has been also interest in understanding what important factors lead some videos to become more popular than others. But few works have studied the temporal aspects of the popularity dynamics using some metrics such as viewcount, ratings and number of comments [3], [10], [11].

In this paper we describe some of the most typical behaviour of the viewcount of videos in YouTube. This allows us to provide in-depth analysis and use models that capture the key properties of the observed popularity dynamics. Our goal is to match observed video viewcounts with one of several dynamic models. To select candidates for these models, we turned to bio-inspired dynamics as we believe that the propagation of a content in YouTube has a strong similarity with the temporal behaviour of an infectious disease, which is a classical topic in mathematical biology [12], [13]. Such models of diseases spread have already been used in order to model the spread of viruses in computer networks [14], [15]. They have been also used in marketing for capturing the life cycle dynamics of a new product [16]. A large number of papers in marketing have shown that product sales life cycle follows an S-curve pattern in which it initially grows at fast rate and it falls off as the limit of the market share is approached [17].

Our contribution can be summarised in the following key points:

(i) We propose six mathematical biology-inspired models and we show that at least 90% of videos in YouTube are associated to one of these six mathematical models with a Mean Error Rate lower than 5%. We further show how to extract the model parameters for each video.

(ii) We study the robustness of these models to the different thematic categories of the video in YouTube and to different values of the peak popularity of the video. We show that the fraction of videos withing a given model is quite robust and shows little dependence on the different thematic categories of the video, except for Education category which has a different behaviour: for this category it seems that the word-of-mouth is the dominate mechanism through which contents are disseminated. The bio-inspired models we selected are further shown to be robust with respect to the peak popularity of the video but the distribution among them is slightly different between those videos that have acquired less than 1000 views and the rest of the videos. In more than 80% of videos in YouTube, the potential population interested in the video increases over time.

(iii) Two of the six models (the *modified negative exponential* and *modified Gompertz* models) cover most of videos in our YouTube dataset (more than 75%). Both models capture the case of immigration process in which the potential population or the ceiling value become dynamic. Further, the *modified negative exponential* characterizes the

dynamic of a non-viral content and it predicts that the accumulated number of views does not contribute to the propagation of the content. This model corresponds to the scenario wherein the content has been broadcasted to a pool of users. On the other side, the *Gompertz model* captures viral videos in which a part of this dynamic is propagated through word-of-mouth.

(iv) We finally use the above classification along with the automatic parameters extraction in order to predict the evolution of videos' viewcount. We consider two scenarios: in the first one we use half of the viewcount curve as a training sequence while in the second one, we take a fixed training sequence that corresponds to the first 50 days in the lifetime of the video. We then compare the predicted curve to the actual one and study the prediction capacity within a given error bound.

The rest of this paper is structured as follows: Section II describes how the dataset used in this work is built, whereas Section III describes the biology-inspired models and their uses. Our data fitting methodology and main results of automatic classification are discussed in § IV and §V, respectively. Section VI concludes the paper.

II. Setting and data

Since we intend to study different types of dynamic evolution of the viewcount in YouTube, we need to collect a huge number of videos which are available to the general public. In this section we describe how we collected the dataset used in this study. On YouTube, a video is accompanied by a set of valuable data as title, upload time, viewcount, related videos. The video webpage also provides some statistics which are available if the content's owner allows it.

YouTube provides two APIs which allow to retrieve some of those data : the YouTube Data API for collecting static data (which are available for every user) and the YouTube Analytics API for seeking video statistics such as dynamics of a content (which are only available for content's owner). Since some data cannot be collected through the APIs, we used a tool named YOUStatAnalyzer [18] in order to collect all valuable data. The collected data are stored in a noSQL database (MongoDB). The noSQL solution has been chosen to allow dynamic insertion of new features for future works. The dataset used for this study contains more than 80000 videos randomly extracted from YouTube and aged between 5 and 2500 days. This dataset contains some static information for each video such as YouTube id, title of the video, name of the author, duration and list of related videos. It also provides the evolution of some metrics (shares, subscribers, watch time and views) in a daily form and in a cumulative form, from the upload day till the date of crawling.

III. Popularity growth patterns

We focus the analysis on viewcount as the main popularity metric of a video. Previous analysis of YouTube showed a strong correlation between viewcount and other metrics as number of comments, favourites and rating. Further, these metrics correlation becomes stronger with popularity [7]. We model the dynamic evolution of viewcount with some mathematical models from the biology. We classify the evolution of viewcount in YouTube using two criteria:

- **Size of the target population:** The first criterion in selecting the model is related to the size of the population that may be potentially interested by the content. We differentiate between models in which the population potentially interested in the content is nearly constant (we call this the "fixed target population property") and those in which it grows in time (inspired by the branching process terminology, we call this "immigration"). The fixed target population property occurs in some video categories in YouTube as news, sport and movies. Indeed, videos in these categories reach quickly the peak of the popularity and then within a short time the diffusion dies out and the viewcount does not further increase.
- **Virality:** The second criterion in the classification concerns the structural virality. A model is said to be viral (or to have the viral property) if contaminated nodes (these are the viewers of a video) have a significant role in the propagation of the video through sharing or embedding. It is non-viral if the propagation of the video essentially relies on broadcast of the video from the source (it is then said to have the broadcast property). In that case, a large fraction of the target population can receive the information directly from the source.

In the following we describe the dynamic models in biology and their uses.

A. Fixed target population

1) Viral content: To describe the viral content with fixed target population, we use the *Logistic model* or the *Gompertz model*. These models have been used in technology forecasting and are referred as "S-shaped" curve. We test them to capture the evolution of viewcount of a video in YouTube since there is a strong similarity between a video posted in YouTube and a new product launched into the marketplace. Indeed, as shown in different problems in marketing, technology product is often growing slowly followed by rapid exponential growth and finally it falls off as limit of market share is approached.

Logistic model: The *Logistic model* (also referred as *Sigmoid model* in this paper) is a common sigmoid function which describes the evolution of viewcount of a video with fixed target population. This is a first order non-linear

differential equation of the form

$$\frac{dS}{dt} = \lambda S(M - S) \tag{1}$$

where S is the viewcount of a video and M is the maximum size of the (potential) population that could access the content. This is a standard equation in epidemiology for describing the evolution of the number of infected individuals under the assumption that all infected nodes have developed an immunity from infection or these infected nodes stay infected and will not be changed to uninfected state. Hence the infection rate is a function of the rate λ and the size of the infected population. A solution to equation (1) is given by

$$S(t) = \frac{M}{1 + (\frac{M - S(0)}{S(0)})e^{-\lambda M t}}$$

This function shows that initial exponential growth is followed by a period in which growth starts to decrease as approaching the maximum size of population.

The S-shape of the *Logistic model* curve is symmetric. But in the context of viewcount, the convex phase and the concave phase could not always be symmetric. For covering these cases we consider the *Gompertz model*.

Gompertz model: A model which deals with the problem of symmetry of the *Logistic model* is given by the following dynamic equation:

$$\frac{dS}{dt} = \lambda S \log(\frac{M}{S}), \tag{2}$$

This model is called *Gompertz model*, and has been also used as diffusion model of product growth. A solution of equation (2) is given by the Gompertz function :

$$S(t) = M \exp\left(-\log(\frac{M}{S(0)}) \exp(-\lambda t)\right),$$

This model is similar to the *Logistic model* but it is not symmetric about the inflection. In general the *Gompertz model* reaches this point early in the growth trend. This behaviour seems to fit well for some YouTube viewcount evolution dynamics.

2) Non-viral content: A non viral content describes the case where users do not contribute on the propagation of the content. This is the case when the time scale of the content diffusion is very large compared to the size of potential population. Hence this dynamic can model the case where contents gain popularity through advertisement and other marketing tools: examples are when advertisement is broadcasted to a very large pool of users of a social network and people access the content at random thereafter. Hence we assume that the evolution dynamic of the content follows the linear differential equation:

$$\frac{dS}{dt} = \lambda(M - S) \tag{3}$$

This model is called the *negative exponential model*. The solution of (3) is given by :

$$S(t) = S(0) + (M - S(0))(1 - e^{-\lambda t})$$

B. Growing population

The assumption that the population is fixed, is often a reasonable approximation when the evolution of the popularity of a content increases quickly and dies out within a short time. But for many cases, this assumption becomes inappropriate when the time before reaching the saturation region is longer. Here we consider the case of immigration process in which the potential population growth and the dynamic of viewcount of a content are intricacy linked. To capture such dependence we consider different growth scenarios that model the viral case and non viral case. In this paper we restrict our study on the case where the target population grows with a fixed speed.

1) Non-viral content : The *linear growth model* $S(t) = S(0) + \lambda t$ describes in a simple way the situation where users do not contribute to propagate the content to other users but the content benefits of the immigration process which gives a linear growth of the viewcount.

Another kind of non-viral curves observed are concave curves (given by the *negative exponential model*) which do not converge to a flat line but become linear at the horizon due to the immigration process influence. Such dynamics could be modelled by modifying solutions of equation (3) where a linear component is added, giving the *modified negative exponential model*:

$$S(t) = S(0) + (M - S(0))(1 - e^{-\lambda t}) + kt$$

where k is the rate of the target population growth.

2) Viral content : Now we address the issue of immigration process in the case of viral contents. In this dynamic the viewcount curve first adopts a viral behaviour (in a S-shaped phase) and then grows linearly.

One candidate solution to describe such a behaviour of viewcount is to add a linear component to the Gompertz function:

$$S(t) = M \exp\left(-\log(\frac{M}{S(0)}) \exp(-\lambda t)\right) + kt$$

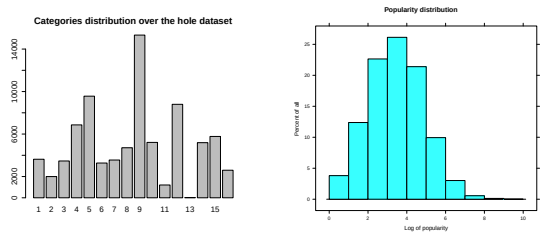
This dynamic, called *modified Gompertz model*, seems to be relevant according to some examples in the dataset.

IV. Dataset and Data fitting

This section describes how we use the models presented in section III in order to classify the YouTube contents in our dataset.

A. Dataset

As described in II, we collected meta-data of more than 80000 videos in a MongoDB database. In addition of the dynamics of viewcount used for modelling, the features we consider for each video are: the age (in number of days), the YouTube category and the popularity (i.e the total number of views at the day of crawling). Fig. 1b shows the distribution of popularity, using logarithmic scales.



(a) YouTube categories dist. (b) Log. dist. of popularity

Figure 1. Some features distributions from the YouTube dataset.

Table I lists the YouTube categories contained within the dataset and Fig. 1a shows their distribution. A summary of age and viewcount values in the dataset is presented in Table II.

Table I
LIST OF ALL YOUTUBE CATEGORIES FOUND INSIDE THE DATASET

1. "Animals"	7. "Games"	13. "Shows"
2. "Autos"	8. "Howto"	14. "Sports"
3. "Comedy"	9. "Music"	15. "Tech"
4. "Education"	10. "News"	16. "Travel"
5. "Entertainment"	11. "Nonprofit"	
6. "Film"	12. "People"	

Table II
SUMMARY OF AGE AND POPULARITY IN THE YOUTUBE DATASET

Age (in days)	**Popularity (number of views)**
Min: 5	Min: 1
1st Qu.: 140	1st Qu.: $2,650.10^2$
Median: 393	Median: $2,728.10^3$
Mean: $610,5$	Mean: $6,091.10^5$
3rd Qu.: 923	3rd Qu.: $2,630.10^4$
Max: 2426	Max: $1,746.10^9$

B. Data fitting

Observations and normalisation: For the data fitting, we only use the cumulative evolution of viewcount as function of time (age). We define a set of observations of a video as follows: $(Y_i, i)_{1\leq i\leq n}$ where Y_i is the viewcount at the day i and n is the number of observations (this is also its age in number of days). In order to avoid some technical issues due to the estimate algorithms, we use normalised observations : $(y_i = \frac{Y_i}{Y_n}, t_i = \frac{i}{n})_{1\leq i\leq n}$.

Parameters estimate methods: We estimate the parameters of the models described in section III using regression algorithms based on the mean squares criterion minimisation. Given a normalised set of observations $(y_i, t_i)_{1\leq i\leq n}$, let S be the expression for one model. The mean squares criterion (MSC) is then given by: $MSC = \sum_i (S(t_i) - y_i)^2$.

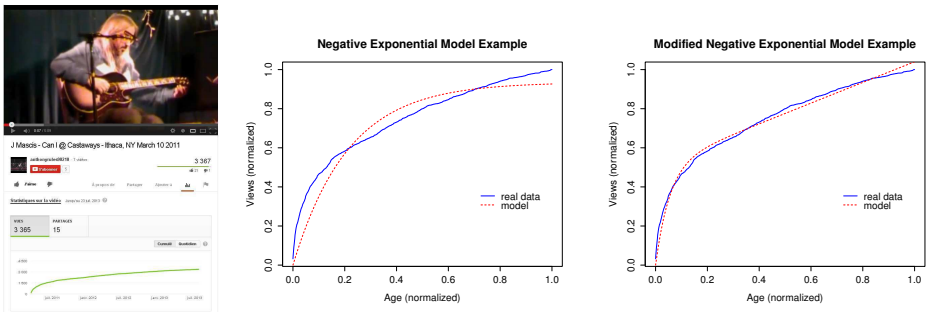
We implemented two algorithms in order to classify the dynamics of any video from YouTube in one of the models presented in section III. The first method is a simple linear regression. It works for videos where the viewcount grows linearly over time t. In that case, the coefficient of determination R gives a measure for the goodness of the fit :

$$R = 1 - \frac{\sum_i (y_i - S(t_i))^2}{\sum_i (y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of $(y_i)_i$. In our experiments, we consider that a linear model is relevant if the value of R satisfies $|R| \geq 0.985$. In the dataset, there are very few video dynamics that match the linear case. The second method is the Levenberg-Marquardt algorithm [19] which is known to be very efficient for the non-linear case. It is an iterative process for estimating parameters of the model through a minimization problem of the MSC. Explicit formulation of models shall be known because the partial derivatives are needed during the iterative process. One drawback of this method, like all other non-linear regression methods, is that the solution could not be global but only a local one. Nevertheless, the Levenberg-Marquardt algorithm suits very well for our models.

Data fitting for Non viral contents: The dynamic described in (3), fits for contents which viewcount curve is concave and it falls off as limit of potential population is approached. In Fig. 2, we show an example where this model is applied. We observe that the estimated curve (dashed line) admits a flat asymptote. But the curve that represents actual data (plain line) seems to follow a line with a non zero slope. This indicates that the potential population may grows over time and is linked with the dynamic of viewcount. In that case, we model the dynamics by the *modified negative exponential model* introduced in subsection III-B1. This model fits better as it is shown in Fig. 2c.

Data fitting for Viral contents: Three models have been considered in the case of viral contents: *Logistic model* and *Gompertz model* for fixed population, and a *modified Gompertz model* for growing population (see III-B2). Fig. 3 is an example where we fit these models to one YouTube content (Fig. 3a). We observe that the S-shape of the *Logistic model* curve is symmetric due to the symmetrical property of sigmoid function (Fig. 3b). However, the convex phase and the concave phase are non symmetric as we can observe in Figure 3a. Hence the *Logistic model* does not fit well. Then, *Gompertz model* and *modified Gompertz model* are fitted to the same YouTube content. The *Gompertz model* (Fig. 3c) fits better than the *Logistic model*, and the *modified*



(a) YouTube content with a concave shape

(b) Viewcount (plain), negative exponential curve (dashed)

(c) Viewcount (plain), modified negative exponential curve (dashed)

Figure 2. From a YouTube content, parameters of the negative exponential model are estimated, then the obtained curve is compared to data (in the centre). The same process is applied to a negative exponential model in which a linear component has been added (on the right side)

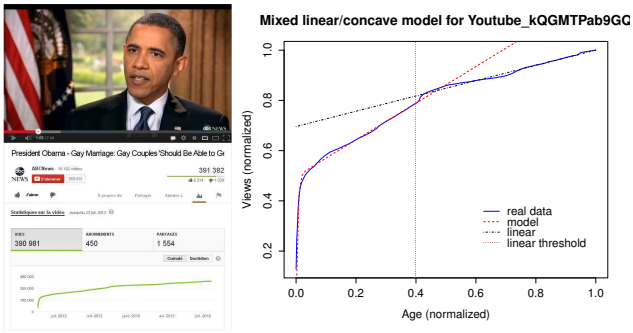
Gompertz model (Fig. 3d) describes better the behaviour of the data at the horizon (immigration phenomena).

Remark 1: We observe two types of behaviour at the horizon: a flat line showing that the limit of the potential population has been reached or an oblique line highlighting the fact that the population continues to grow. However, both cases indicate a linear phase. If this phase is large (i.e. lasts longer regarding the age of a video), then it might smoothen the error done on the early phase. Thus, the viral property may not be well captured. In order to address this timescale issue, we propose to use two phases for data fitting. Fig. 4 gives an example where this method is applied for a YouTube content (Fig. 4a). Indeed, given a set of observations $(y_i, t_i)_{1 \leq i \leq n}$, a first phase consists to find out a linear behaviour from a time $t = t_k, k \in 1, .., n$. The idea is to find k in order to have a good regression line for the subset $(y_i, t_i)_{k \leq i \leq n}$[1]. In Fig. 4b, the time t_k is around 0.4 from which the evolution of viewcount can be well modelled by a linear model. In the second phase, the evolution of viewcount is estimated using data fitting presented in previous sections. Fig. 4b illustrates this phase in which the viewcount curve on the left side of $t_k = 0.4$, is associated to the *modified negative exponential model*. Actually, this method is not implemented yet in our classification presented in next section. The gain of this technique is addressed in the technical report [20].

V. Automatic classification

The main goal of our work is to provide a system that can automatically classify YouTube contents by associating one model to one content. For each content, two issues have to be managed: first, we evaluate each model in order to know which models are good candidates and we compare the selected candidates in order to determine which one is the best.

[1]In the technical report [20], we describe in detail the algorithm that we used to compute the time t_k in which observations $(y_i, t_i)_{k \leq i \leq n}$ show a linear behaviour



(a) YouTube video

(b) Mixed linear/non linear fitting

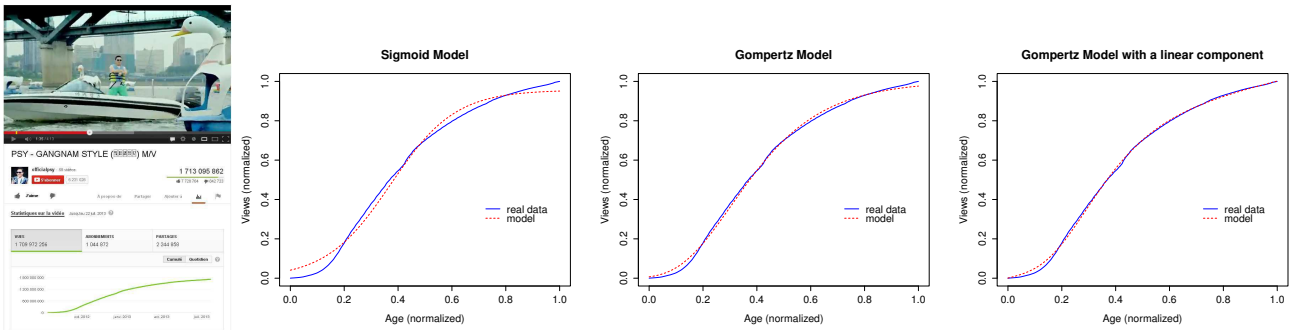
Figure 4. From a YouTube content (a), a mixed procedure is made to estimate a linear model and a non linear model on two subset of data (b)

Let us consider first the question of evaluating each model. As explained in section IV-B, we perform parameters estimate based on the least squares criterion minimisation. Define the mean error rate (MER) by

$$MER = \frac{1}{n} \sum_i \frac{|S(t_i) - y_i|}{y_i + 1}$$

MER criterion is the mean error rate done by the model regarding the observations. For example, if $MER \leq 0,05$, it can be said that on average, the estimate error is lower than 5% of the observed value. The choice of the $(y_i + 1)$ terms in the denominator aims at smoothing large values of MER generated by some very small values of y_i. This issue is disscused in more detail in [20]. With this criterion we fix a threshold beyond which one model would be considered as unreliable. In order to compare models with MER lower than this threshold, we introduce a criterion of quality discussed in [21]. To formulate this criterion we first define the degree of freedom of a model by $df = n - p$ where p is the number of parameters of the model. The criterion of quality, named "goodness of fit" (GoF) is then given by:

$$GoF = \frac{1}{(df)} MSC$$



(a) YouTube video (b) Logistic model fitting (c) Gompertz model fitting (d) Modified Gompertz model fitting

Figure 3. From a YouTube video with a S-shaped viewcount curve 3a, fit of the Logistic model is done in 3b. The estimated curve (dashed) is compared with the actual normalised viewcount curve (plain). The same is done with the Gompertz model in 3c and finally with the modified Gompertz model in 3d.

Table V
CONFIDENCE INTERVALS FOR MODELS DISTRIBUTION PROPORTION

Model/sampling	1000	10000	$Wholedataset(81657)$
Exponential	$(0.0573 - 0.0722 - 0.0904)$	$(0.0668 - 0.0717 - 0.077)$	$(0.07 - 0.0718 - 0.0736)$
Modified Exponential	$(0.339 - 0.3688 - 0.3997)$	$(0.3584 - 0.3679 - 0.3774)$	$(0.3648 - 0.3681 - 0.3714)$
Sigmoid	$(0.0214 - 0.0308 - 0.0441)$	$(0.0273 - 0.0306 - 0.0342)$	$(0.0292 - 0.0304 - 0.0316)$
Modified Sigmoid	$(0.0222 - 0.0318 - 0.0452)$	$(0.0277 - 0.031 - 0.0347)$	$(0.0298 - 0.0309 - 0.0322)$
Gompertz	$(0.0153 - 0.0234 - 0.0353)$	$(0.0206 - 0.0234 - 0.0267)$	$(0.0226 - 0.0236 - 0.0247)$
Modified Gompertz	$(0.331 - 0.3607 - 0.3914)$	$(0.3535 - 0.3628 - 0.3723)$	$(0.3595 - 0.3627 - 0.3661)$

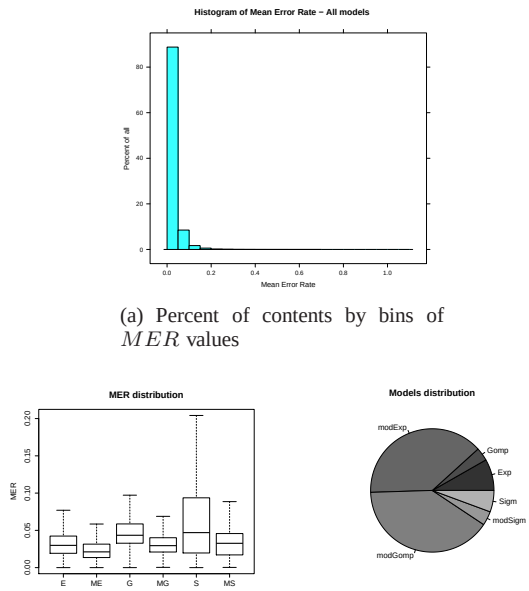
Table III
GOODNESS OF FIT FOR MODELS FROM FIG. 2

Model	MCS	MER	GoF
Neg. exponential	3.558	0.074	0.004
Modified neg. exponential	0.453	0.027	$4.98.10^{-4}$

Table IV
GOODNESS OF FIT FOR MODELS FROM FIG. 3

Model	MCS	MER	GoF
Sigmoid	0.480	0.021	10^{-3}
Gompertz	0.092	0.018	$1.846.10^{-4}$
Modified Gompertz	0.033	0.008	$8.831.10^{-5}$

The model which has the smallest GoF will be considered as the best one. In Table III and Table IV, we list values of MSC, MER and GoF for models used respectively in Fig. 2 and Fig. 3. In the example from Fig. 2, with a threshold of MER fixed at 0.075, both *negative exponential model* and *modified negative exponential model* are relevant. With a GoF of value $4.98.10^{-4}$, *modified negative exponential model* is the one that fits best. In the case of YouTube content depicted in Fig. 3, if the threshold of MER is fixed at 0.02, *Sigmoid model* (i.e *Logistic model*) is not reliable whereas *Gompertz model* and *modified Gompertz model* respect the threshold constraint. According to the value of GoF, *modified Gompertz model* is the best with $GoF = 8.831.10^{-5}$. Further, the issue of fixing a value for the MER threshold is crucial to rely on an acceptable filter for several videos. Next step is to associate each video in our dataset to a mathematical model by using MER and GoF criteria. Our classification using six mathematical models, shows that 90% of videos are associated to a model with a MER lower than 0.05 (see Fig. 5a). A mean error of 5% seems reasonable to consider a reliable fitting. Note



(a) Percent of contents by bins of MER values

(b) MER distribution for each model

(c) Models distribution after classification over the whole dataset

Figure 5. Sample analysis of an automatic classification for dissemination processes in YouTube

that if the threshold of MER is fixed at 0.1, more than

97% of the videos correspond to one of the models. There is at most 2% of the videos for which the association to one of our models gives a high error rate (let say more than 10%). Fig. 6 illustrates one example of such a video. The association is unreliable due to the many sharp changes of the behaviour. Indeed, it seems that the models are unable to capture the effect of multiple peaks in viewcount evolution. The investigation of this type of videos will be studied in future works.

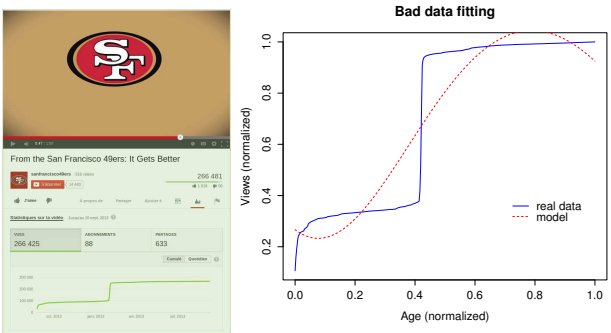


Figure 6. One bad fitting example.

In Fig. 5b, we show the MER distribution for each model. We also introduce a new model, named *modified sigmoid model*, given by the *Logistic model* with a linear component added (as done in section III-B2 with the *modified Gompertz model*). In this figure, the letters E, ME, G, MG, S and MS are respectively for *negative exponential*, *modified negative exponential*, *Gompertz*, *modified Gompert*, *sigmoid* (or *Logistic*) and *modified sigmoid*. It appears that the *ME* and the *MG* give fitting with less error compared to other models. They can be referred as the most reliable models for our dataset. Further, these models cover almost 75% of videos in our dataset (see Fig. 5b). Both models represent the case of immigration process in which the potential population may grow over time. Given this finding, we conclude that most videos in Youtube are still attracting viewers even after a long period. Moreover, it appears that there is a balance between viral and non viral contents.

In order to assess the evidence provided by our dataset on models distribution, we provide in Table V 95% confidence interval for different sample sizes involved in the study. This table indicates that the whole dataset leads to very good precision on models distribution. Furthermore, a sampling with 10000 videos still gives an accurate estimate of this proportion.

Now it is natural to ask whether the distribution of our classification is still the same with respect to main categories in YouTube and popularity of a video. To address these questions, we make the classification in each category as shown in Fig. 1a, and focus on four main categories: Music (over 14000 videos), Entertainment (over 8500 videos), People (around 7500 videos) and Education (almost 6000 videos). In general, the models distribution in each category is quite robust and shows little dependence on the different thematic categories of the video, except for Education category where there is more than 50% of the videos that belong to the *modified Gompertz model*. Moreover, viral models cover almost 75% of the videos. This category seems that the word-of-mouth is the dominate mechanism through which contents are disseminated.

We also analyse the models distribution by considering different classes of popularity. According to the distribution depicted in Fig. 1b, we define seven classes of popularity listed in Table VI. We show the models distribution for each popularity class in Table VII.

Table VI
POPULARITY CLASSES

Popularity class	Total number of views V
Extremely unpopular (EUP)	$0 \leqslant V < 10$
Very unpopular (VUP)	$10 \leqslant V < 100$
Unpopular (UP)	$100 \leqslant V < 1000$
Not so popular (NSP)	$1000 \leqslant V < 10^4$
Popular (P)	$10^4 \leqslant V < 10^5$
Very popular (VP)	$10^5 \leqslant V < 10^6$
Extremely popular (EP)	$10^6 \leqslant V$

Table VII
MODELS DISTRIBUTION BY POPULARITY CLASS (IN %)

Model	EUP	VUP	UP	NSP	P	VP	EP
Exp	11.4	12.2	8.4	8	6.8	6.2	5.7
Gomp	1.6	2.8	1.8	2.5	3.6	3.2	2.1
ModExp	11.6	54.5	48.9	35.2	35.1	42.3	47.5
ModGomp	2.5	19.4	34.7	48.7	49.3	44.3	42.8
ModSigm	1.8	4.3	4.3	3.6	2.9	2.3	0.8
Sigm	70.7	6.6	1.5	1.7	2	1.3	0.8

We observe that the distribution varies according to the classes of popularity. First of all, the *sigmoid model* dominates the *extremely unpopular* videos (constituted by videos of less than 10 views). These results are not reliable due to the few different values of the viewcount for these videos. *Popular* videos and *not so popular* videos can be grouped in terms of models distribution with around 50% for *modified Gompertz model* and 35% for *modified exponential model*. We can also group *very popular* videos and *extremely popular* videos for which distribution is slightly equivalent to the whole dataset distribution (see Fig. 5c). *Very unpopular* and *unpopular* videos exhibit *modified exponential model* around 50% of the cases. The *modified Gompertz model* represents less than 20% in *very unpopular* videos whereas it covers almost 35% of the videos in *unpopular* videos. More investigation are reported in [20], in particular for models distribution as a function of popularity or category, and also for the prediction which is briefly introduced in the following.

Classification models for prediction

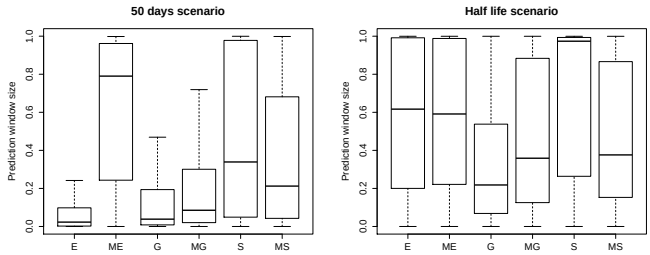
In this section, we illustrate a mechanism for predicting the future evolution of viewcount of a video. In particular, we propose a simple model that predicts the evolution of viewcount from a given date t_f till a target date t_p with $t_p > t_f$. We call a prediction window T the difference between t_p and t_f. This prediction is based on the early historical information of a video which is given by a set of observations $(y_i, t_i)_{1 \leq i \leq f}$ till time t_f where f is number of observations[2]. Combining these information with our classification models, the evolution of viewcount is estimated using data fitting in order to select a mathematical model. Using our datasets, we evaluate the maximum size of prediction window with at most 5% mean error, i.e

$$T_{max} = \max\{t_p - t_f | \frac{1}{p-f} \sum_{i=f+1}^{p} \frac{|S_{t_f}(t_i) - y_i|}{(y_i + 1)} \leq 0.05\}$$

where S_{t_f} is the selected mathematical model.

Table VIII
MEAN AND VARIANCE OF PREDICTION WINDOW SIZE IN THE HALF LIFE SCENARIO

Model	mean	var	number of videos
E	0.5833692	0.1504571	4132
ME	0.576914	0.1415303	25281
G	0.3435265	0.1160928	683
MG	0.4596889	0.1360676	19349
S	0.6765688	0.1544357	1030
MS	0.4625144	0.1280855	1659



(a) 50 days classification (b) Half life classification

Figure 7. Prediction window size according to models type

We test our prediction for the scenario in which t_f corresponds to half life cycle. Let $\Delta T = \frac{T_{max}}{t_n - t_f}$ where $t_n - t_f$ is the remaining time of life cycle of a video from t_f. Note that ΔT is bounded by 1. Fig. 7b depicts the mean and variance of ΔT for each identified model. Table VIII precises values of mean and variance for each model as well as the number of videos classified in the different models. Our results show that our prediction is very powerful and most models provide a prediction window that long enough within an error bound at 5%. Further we observe that our scheme can perfectly predict the evolution of viewcount till the half of the remaining time of life cycle from the time of prediction.

We tested here the prediction based on a learning sequence that was half the lifetime of each video in the dataset. This allows the prediction to rely on the same amount of data independently of the real duration of the video. We next compare this to the case in which, in contrast, the learning sequence has a fixed duration of 50 days. We note that 50 days represent much less than half the lifetime for most videos in the data set and therefore the prediction is less accurate. The corresponding results of this scenario are depicted in Table IX and Fig. 7b. In spite of this problem we get similar results of the average prediction window for models *modified Gompertz* and *sigmoid* (*Logistic*).

Table IX
MEAN AND VARIANCE OF PREDICTION WINDOW SIZE IN THE 50 DAYS SCENARIO

Model	mean	var	number of videos
E	0.1240775	0.06050271	3401
ME	0.6159881	0.1384898	21788
G	0.1861161	0.09028605	687
MG	0.2314134	0.09438778	13821
S	0.4750911	0.1083194	1137
MS	0.2082774	0.1798454	1561

VI. CONCLUSION AND FUTURE WORK

In this work we provided an automatic way for classifying the dynamic evolution of viewcount of videos into six mathematical models that are inspired from the biology. We showed that most of videos in our datasets are associated to a model with a MER lower than 0.05, which indicates the prefect matching between the mathematical model and observed datasets. Using our classification, we provided several findings. First, we were able to characterise the key properties of videos as virality and potential population growth. Second, two of the six models often appeared as the best candidates to fit well with datasets and both models expect that most videos experienced an immigration process in which the potential population grows over time. Third, we developed a rigorous model that allow us to predict the evolution of viewcount during a window time.

Based in this work, we identify many directions that we expect to study in the future work. (i) Conducting larger scale datasets collection, because some characteristics might be perceived only when the datasets is fairly large. (ii) Refine the distribution of models based on uploader characteristics (e.g uploader network, followers, audience of previous videos, etc) which can be useful for prediction of future popularity. (iii) Investigating how the size of prediction window evolves as function of the time of prediction. (iv) Another possible direction is to investigate

[2]The datasets used for prediction contains videos with at least 50 days old.

how data from google Trends related to some videos, affect our classification as well as the models distribution.

ACKNOWLEDGMENT

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.

REFERENCES

[1] "Comscore. more than 200 billion online videos viewed globally in october," http://www.comscore.com/Press_Events/Press_Releases/2011/12/, December 2011.

[2] "Comscore danpiech. online video by the numbers," http://www.comscore.com, July 2011.

[3] 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.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] G. Szabo and B. A. Huberman, "Predicting the Popularity of Online Content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, aug 2010.

[10] X. Cheng, C. Dale, and J. Lui, "Statistics and social network of youtube videos," *In Proc. International Workshop on Quality of Service (IWQoS) The Netherlands*, p. 229 238, June, 2008.

[11] S. Mitra, M. Agrawal, A. Yadav, N. Carlsson, D. Eager, and A. Mahanti, "Characterizing web-based video sharing workloads," *ACM Transactions on the Web*, vol. 2, no. 8, pp. 8 – 27, 2011.

[12] N. Bailey, *The Mathematical Theory of Infectious Diseases and its Applications*. London: Griffin, 1975.

[13] L. A. Meyers, "Contact network epidemiology: Bond percolation applied to infectious disease prediction and control," *Bull. AMS*, vol. 44, no. 1, pp. 63–86, 2007.

[14] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos, "Epidemic thresholds in real networks," *ACM Trans. Inf. Syst. Secur.*, vol. 10, no. 4, pp. 1:1–1:26, jan 2008.

[15] A. Ganesh, L. Massoulie, and D. Towsley, "The effect of network topology on the spread of epidemics," in *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, vol. 2, Miami, FL, USA, March 2005, pp. 1455–1466.

[16] F. M. Bass, "The relationship between diffusion rates, experience curves, and demand elasticities for consumer durable technological innovations," *The Journal of Business*, vol. 53, no. 3, pp. pp. 51–67, 1980.

[17] V. Mahajan, E. Muller, and Y. Wind, *New-Product Diffusion Models*, ser. International Series in Quantitative Marketing. Springer, 2000.

[18] M. Zeni, D. Miorandi, and F. De Pellegrini, "YOUStatAnalyzer: a tool for analysing the dynamics of YouTube content popularity," in *Proc. 7th International Conference on Performance Evaluation Methodologies and Tools (Valuetools, Torino, Italy, December 2013)*, Torino, Italy, 2013.

[19] D. W. Marquardt, "An algorithm for mean-squares estimation of nonlinear parameters," *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, jun 1963.

[20] C. Richier, E. Altman, R. Elazouzi, T. Jimenez, G. Linares, and Y. Portilla, "Modelling view-count dynamics in youtube," *Technical Report url:http://arxiv.org/abs/1404.2570*, 2014. [Online]. Available: https://arxiv.org/pdf/953787

[21] W. E. Deming, "The chi-test and curve fitting," *Journal of the American Statistical Association*, vol. 29, no. 188, pp. 372–382, Dec 1934.