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
†INRIA, B.P 93, 06902 Sophia Antipolis Cedex, FRANCE

Email: firstname.lastname@univ-avignon.fr
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 video 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 that 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.

- **Viralilty**: 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
\[
\frac{dS}{dt} = \lambda S(M - S)
\] (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 \( \lambda \) and the size of the infected population. A solution to equation (1) is given by
\[
S(t) = \frac{M}{1 + (\frac{M - S(0)}{S(0)}) e^{-\lambda Mt}}
\]
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:
\[
\frac{dS}{dt} = \lambda S \log \left( \frac{M}{S} \right),
\] (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:
\[
S(t) = M \exp \left( -\log \left( \frac{M}{S(0)} \right) \exp(-\lambda t) \right),
\]
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:
\[
\frac{dS}{dt} = \lambda (M - S)
\] (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:
\[
S(t) = M \exp \left( -\log \left( \frac{M}{S(0)} \right) \exp(-\lambda t) \right) + kt
\]
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.

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>
<thead>
<tr>
<th>Table I</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LIST OF ALL YOUTUBE CATEGORIES FOUND INSIDE THE DATASET</strong></td>
</tr>
<tr>
<td>1. &quot;Animals&quot;</td>
</tr>
<tr>
<td>5. &quot;Entertainment&quot;</td>
</tr>
<tr>
<td>6. &quot;Film&quot;</td>
</tr>
</tbody>
</table>

Table II |
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SUMMARY OF AGE AND POPULARITY IN THE YOUTUBE DATASET</strong></td>
</tr>
<tr>
<td>Age (in days)</td>
</tr>
<tr>
<td>Min: 5</td>
</tr>
<tr>
<td>1st Qu.: 140</td>
</tr>
<tr>
<td>Median: 395</td>
</tr>
<tr>
<td>Mean: 610.5</td>
</tr>
<tr>
<td>3rd Qu.: 922</td>
</tr>
<tr>
<td>Max: 2426</td>
</tr>
</tbody>
</table>

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: \((Y_i, t_i)_{1 \leq i \leq n}\) 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: \((\bar{y}_i = \frac{y_i}{\bar{N}}, \bar{t}_i = \frac{t_i}{\bar{N}})_{1 \leq i \leq n}\).

**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 \((y_i)\). 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
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.

(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.

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 \( (M_{ER}) \) by

\[
M_{ER} = \frac{1}{n} \sum_{i=1}^{n} \frac{|S(t_i) - y_i|}{y_i + 1}
\]

\( M_{ER} \) criterion is the mean error rate done by the model regarding the observations. For example, if \( M_{ER} \leq 0.05 \), 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 \( M_{ER} \) generated by some very small values of \( y_i \). This issue is discussed 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 \( M_{ER} \) 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:

\[
GoF = \frac{1}{(df)} \cdot M_{SC}
\]
wholedataset threshold of MER is the one that fits best. In the case of YouTube content depicted in Fig. 3, if the threshold of MER is respected, the threshold constraint. According to the GoF value of 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

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

![YouTube video](image)

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>
<thead>
<tr>
<th>Model/sampling</th>
<th>1000</th>
<th>10000</th>
<th>whole dataset (81657)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>(0.0573 – 0.0722 – 0.0904)</td>
<td>(0.0668 – 0.0717 – 0.077)</td>
<td>(0.07 – 0.0718 – 0.0736)</td>
</tr>
<tr>
<td>Modified Exponential</td>
<td>(0.329 – 0.3688 – 0.3997)</td>
<td>(0.3584 – 0.3679 – 0.3774)</td>
<td>(0.3648 – 0.3691 – 0.3714)</td>
</tr>
<tr>
<td>Modified Sigmoid</td>
<td>(0.0214 – 0.0308 – 0.0441)</td>
<td>(0.0273 – 0.0306 – 0.0342)</td>
<td>(0.0292 – 0.0304 – 0.0316)</td>
</tr>
<tr>
<td>Gompertz</td>
<td>(0.0222 – 0.0318 – 0.0452)</td>
<td>(0.0277 – 0.031 – 0.0347)</td>
<td>(0.0298 – 0.0309 – 0.0322)</td>
</tr>
<tr>
<td>Modified Gompertz</td>
<td>(0.0153 – 0.0234 – 0.0353)</td>
<td>(0.0206 – 0.0234 – 0.0267)</td>
<td>(0.0226 – 0.0236 – 0.0247)</td>
</tr>
<tr>
<td></td>
<td>(0.331 – 0.3607 – 0.3914)</td>
<td>(0.3535 – 0.3628 – 0.3723)</td>
<td>(0.3595 – 0.3627 – 0.3661)</td>
</tr>
</tbody>
</table>

![Histogram of Mean Error Rate - All models](image)

(a) Percent of contents by bins of MER values

![MER distribution](image)

(b) MER distribution for each model

![Models distribution](image)

(c) Models distribution after classification over the whole dataset

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

![Table III - Goodness of fit for models from Fig. 2](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>MCS</th>
<th>MER</th>
<th>GoF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neg. exponential</td>
<td>3.558</td>
<td>0.074</td>
<td>0.004</td>
</tr>
<tr>
<td>Modified neg. exponential</td>
<td>0.453</td>
<td>0.027</td>
<td>4.98 × 10^{-4}</td>
</tr>
</tbody>
</table>

![Table IV - Goodness of fit for models from Fig. 3](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>MCS</th>
<th>MER</th>
<th>GoF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sigmoid</td>
<td>0.480</td>
<td>0.021</td>
<td>10^{-4}</td>
</tr>
<tr>
<td>Gompertz</td>
<td>0.002</td>
<td>0.018</td>
<td>1.840 × 10^{-4}</td>
</tr>
<tr>
<td>Modified Gompertz</td>
<td>0.003</td>
<td>0.008</td>
<td>9.831 × 10^{-5}</td>
</tr>
</tbody>
</table>
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.

![Bad data fitting](image)

Figure 6. One bad fitting example.

In Fig. 5b, we show the $M.E.R$ 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 Gompertz, 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](image)

**Table VI**

<table>
<thead>
<tr>
<th>Popularity class</th>
<th>Total number of views $t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely unpopular (EUP)</td>
<td>$0 \leq V &lt; 10$</td>
</tr>
<tr>
<td>Very unpopular (VUP)</td>
<td>$10 \leq V &lt; 100$</td>
</tr>
<tr>
<td>Unpopular (UP)</td>
<td>$100 \leq V &lt; 1000$</td>
</tr>
<tr>
<td>Not so popular (NSP)</td>
<td>$1000 \leq V &lt; 10^4$</td>
</tr>
<tr>
<td>Popular (P)</td>
<td>$10^4 \leq V &lt; 10^6$</td>
</tr>
<tr>
<td>Very popular (VP)</td>
<td>$10^6 \leq V &lt; 10^9$</td>
</tr>
<tr>
<td>Extremely popular (EP)</td>
<td>$V \geq 10^9$</td>
</tr>
</tbody>
</table>

![Table VII](image)

**Table VII**

<table>
<thead>
<tr>
<th>Model</th>
<th>EUP</th>
<th>VUP</th>
<th>UP</th>
<th>NSP</th>
<th>P</th>
<th>VP</th>
<th>EP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp</td>
<td>11.4</td>
<td>12.2</td>
<td>8.4</td>
<td>8</td>
<td>6.8</td>
<td>6.2</td>
<td>3.7</td>
</tr>
<tr>
<td>Gomp</td>
<td>1.6</td>
<td>2.8</td>
<td>1.8</td>
<td>2.5</td>
<td>3.6</td>
<td>3.2</td>
<td>2.1</td>
</tr>
<tr>
<td>ModExp</td>
<td>11.6</td>
<td>34.4</td>
<td>48.9</td>
<td>35.4</td>
<td>45.4</td>
<td>42.3</td>
<td>17.8</td>
</tr>
<tr>
<td>ModGomp</td>
<td>2.5</td>
<td>19.4</td>
<td>34.7</td>
<td>48.7</td>
<td>49.3</td>
<td>44.3</td>
<td>42.8</td>
</tr>
<tr>
<td>ModSigm</td>
<td>1.8</td>
<td>4.3</td>
<td>4.3</td>
<td>3.6</td>
<td>2.9</td>
<td>2.3</td>
<td>0.8</td>
</tr>
<tr>
<td>Origin</td>
<td>70.7</td>
<td>6.0</td>
<td>1.5</td>
<td>1.7</td>
<td>2</td>
<td>1.3</td>
<td>0.8</td>
</tr>
</tbody>
</table>

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. 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.

\[
T_{max} = \max \{ t_p - t_f \left \lfloor \frac{1}{p-f} \sum_{i=f+1}^{p} |S_{t_i}(t_i) - y_i| (y_i+1) \right \rfloor \leq 0.05 \}
\]

where \( S_{t_i} \) is the selected mathematical model.

<table>
<thead>
<tr>
<th>Model</th>
<th>mean</th>
<th>var</th>
<th>number of videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>0.5834692</td>
<td>0.1504571</td>
<td>4132</td>
</tr>
<tr>
<td>ME</td>
<td>0.5760914</td>
<td>0.1413606</td>
<td>25284</td>
</tr>
<tr>
<td>G</td>
<td>0.3435265</td>
<td>0.1106928</td>
<td>683</td>
</tr>
<tr>
<td>MG</td>
<td>0.4596989</td>
<td>0.1360676</td>
<td>19349</td>
</tr>
<tr>
<td>S</td>
<td>0.6765688</td>
<td>0.1544357</td>
<td>1090</td>
</tr>
<tr>
<td>MS</td>
<td>0.4625144</td>
<td>0.1280855</td>
<td>1659</td>
</tr>
</tbody>
</table>

![50 days classification](image1)

![Half life classification](image2)

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 \( \Delta T = \frac{T_{max}}{t_f} \) where \( t_n - t_f \) is the remaining time of life cycle of a video from \( t_f \). Note that \( \Delta T \) is bounded by 1. Fig. 7b depicts the mean and variance of \( \Delta T \) 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>
<thead>
<tr>
<th>Model</th>
<th>mean</th>
<th>var</th>
<th>number of videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>0.1240915</td>
<td>0.0890423</td>
<td>7491</td>
</tr>
<tr>
<td>ME</td>
<td>0.6150981</td>
<td>0.1384086</td>
<td>21788</td>
</tr>
<tr>
<td>G</td>
<td>0.1861161</td>
<td>0.0902860</td>
<td>687</td>
</tr>
<tr>
<td>MG</td>
<td>0.2314134</td>
<td>0.0943877</td>
<td>13921</td>
</tr>
<tr>
<td>S</td>
<td>0.4750911</td>
<td>0.1081914</td>
<td>1137</td>
</tr>
<tr>
<td>MS</td>
<td>0.2062774</td>
<td>0.1198454</td>
<td>1561</td>
</tr>
</tbody>
</table>

**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
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


