Information Sciences 369 (2016) 585–613

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Predicting the popularity growth of online content: Model and algorithm

Ilias N. Lymperopoulos

Department of Management Science and Technology, Athens University of Economics and Business, 47a Evelpidon Str., Athens, 11362, Greece

ARTICLE INFO

Article history:
Received 9 June 2015
Revised 25 June 2016
Accepted 17 July 2016
Available online 18 July 2016

Keywords:
Popularity prediction
Information diffusion
Information cascade
Social network
Online content adoption
Popularity growth model

ABSTRACT

The evolution of the popularity of online content is analyzed, and two characteristic patterns pertaining to linear and non-linear growth periods are detected. While the former characterizes the propagation of online content through a dynamical process in a state of statistical equilibrium, the latter appears when this state is perturbed by exogenous intervention events. Such episodes increase the susceptibility of higher threshold individuals who opportunistically adopt the propagating content. To capture the dynamics of both diffusion modes, the popularity of online content is modeled by interlacing linear and non-linear growth terms, reduced to 1st-degree polynomial and logistic functions corresponding respectively to stationary and non-stationary adoption phases. The precise fit of the model to empirical popularity patterns verifies its suitability as prediction tool. The proposed model is employed to generate forecasts about the popularity of online content through extrapolation. Highly accurate prediction results surpassing existing methods in terms of precision and predictive capacity are demonstrated. The prediction method is formulated into an algorithm, applicable to real time forecasting of the popularity of online content without training, using minimal, macroscopic, publicly available information.

1. Introduction

As online social networks become ubiquitous and develop into the primary venue for the dissemination of information generated by individuals, businesses and organizations, the prediction of the popularity of online content is increasingly important for reputation management, growth of business opportunities, and effective communication. Reliable forecasts provide valuable insights into the content quality and its potential reach, thereby either corroborating the online communication efficiency, or providing timely signs for remedial action. Undoubtedly, forecasting the diffusion of online content is vital to a successful online presence, but nonetheless it begs the question: Is it possible to make predictions when a multitude of unforeseen factors affect the adoption dynamics?

The diffusion of online content is a dynamical process exhibiting time-varying behavior, and therefore prediction is inherently interlinked with the temporal characteristics of the popularity evolution. To formulate a prediction method, we follow a three-stage approach comprising the detection of universal popularity growth patterns, their modeling as a function of time, and the use of the derived model for generating popularity forecasts. The fitting of the model to real popularity patterns is essential in order to provide sufficient evidence of its adequacy to describe the popularity growth process, thereby

E-mail address: eliasliber@aueb.gr, eliasliber@hotmail.com

http://dx.doi.org/10.1016/j.ins.2016.07.043
0020-0255/

justifying its use as a prediction tool. The estimation of the model parameter values from available information makes it possible to predict the evolution of the online content popularity through forward extrapolation.

The implementation of the foregoing approach starts with the analysis of empirical data. In this regard, we examined time-series describing the popularity growth of Twitter hashtags, Memetracker phrases, and Flickr photos with a view to detecting common patterns, descriptive of universal principles in the underlying adoption dynamics of online content. The time-series are sample realizations from an infinite number of instances that could be potentially generated by an online content diffusion process. However, by inspecting samples from different online domains we can infer the statistical properties of the online content adoption dynamics that are invariant among all possible realizations, thereby fulfilling a primary objective of statistical investigation.

The data analysis showed that the popularity of online content evolves through two operating modes. The first one pertains to the steady-state of a stochastic dynamical process, whereby the adoption rate follows a stationary pattern, indicative of the statistical equilibrium state [6,15] of a content diffusion process. In this state the probability laws driving the content propagation remain invariant, thereby rendering the average adoption rate constant and independent of time, though small fluctuations still appear in the diffusion pattern. During this state the popularity growth of online content follows a linear pattern. Such an operating mode emerges when an online social network functions as a closed system, where the content adoption is driven by the activity of a homogeneous population of users who search for information and share links with their neighbors. Such events are mostly independent and result in trivial adoption cascades causing minor fluctuations in a stable adoption rate characterizing a content diffusion process in a state of statistical equilibrium. Stationary adoption patterns represent an enduring operating state and dominant condition in the online content diffusion for considerable stretches of time, especially for persistent online content. The second mode in the online content adoption process results from perturbations caused by the contact between an online social network and its surroundings. This contact takes place through intervention events significantly affecting the content adoption pattern which acquires the form of the individuals' response to information flows, such as news, advertising, and high-profile events. The exogenous excitation changes the susceptibility of a subpopulation of users with heterogeneous activation thresholds, thereby generating conditions favoring the opportunistic adoption of the propagating online content. The analysis of empirical data showed that the popularity evolution patterns generated by intervention events constitute a single class described by the logistic growth dynamics [30,42]. This means that the adoption of online content during non-stationary diffusion periods is analogous to the adoption of innovations [5,34,40], threshold phenomena in collective behavior [12,17], and other self-limiting growth processes pertaining to biology, ecology, epidemics and demography [30]. Indicative examples of the popularity evolution of online content are depicted in Fig. 1. The universality of the illustrated patterns among all types of online content allows the generalization of the popularity growth dynamics, which can be expressed in the form of a model comprising intermingled linear and logistic growth phases much like an input-dependent renewal process consisting of hold and jump periods [9]. By modeling these patterns and estimating the values of the model parameters from real data, we demonstrate that we can accurately predict the popularity growth of online content regardless of the time resolution of the available information, the type, popularity level and persistence of content, as well as the length of the prediction horizon. The proposed prediction method is differentiated from existing approaches in many respects. In particular:

- It provides precise estimations of the popularity growth of online content in linear and non-linear adoption periods using minimal, macroscopic, real time information. To the best of our knowledge such forecasts are not generated by existing methods.
- It is based on a prediction model accurately fitting time-varying popularity patterns. Existing methods do not provide evidence of fitting dynamic adoption patterns.
- It is universal and applicable to the popularity prediction of any type of online content, irrespective of the content popularity and persistence.
- It can be applied both to the prediction of the number of adopters, and to the prediction of the activity level in relation to a particular piece of online content.
- It does not require training and information about the content attributes, the users' characteristics, the social network structure, and the early adoption patterns.
- The modeling of popularity growth patterns is independent of the time resolution of the available data, thus enabling the generation of accurate predictions using various timescales.

The simplicity, generality, flexibility and accuracy of the proposed method are its major strengths contributing to an effective solution to the online content popularity prediction problem in real time with limited data. The rest of the paper is organized as follows. In Section 2 we review studies relevant to the popularity prediction of online content. In Section 3 we utilize the phenomenological observations in order to construct an equation-based model fitting the popularity evolution of online content. In Section 4 we discuss how the proposed model can be incorporated into a method aiming at generating popularity forecasts using information about the adoption rate and its acceleration. In Section 5 we provide a series of experiments confirming: (a) The effectiveness of the proposed method; (b) the substantial improvement on the performance of benchmark models; and (c) the highly precise fitting of the equation-based predictive model to real popularity patterns. In Section 6, we provide an algorithmic implementation of the proposed method addressing the popularity prediction of online content in real time. Finally, in Section 7 we discuss the findings of this study and present the concluding remarks.

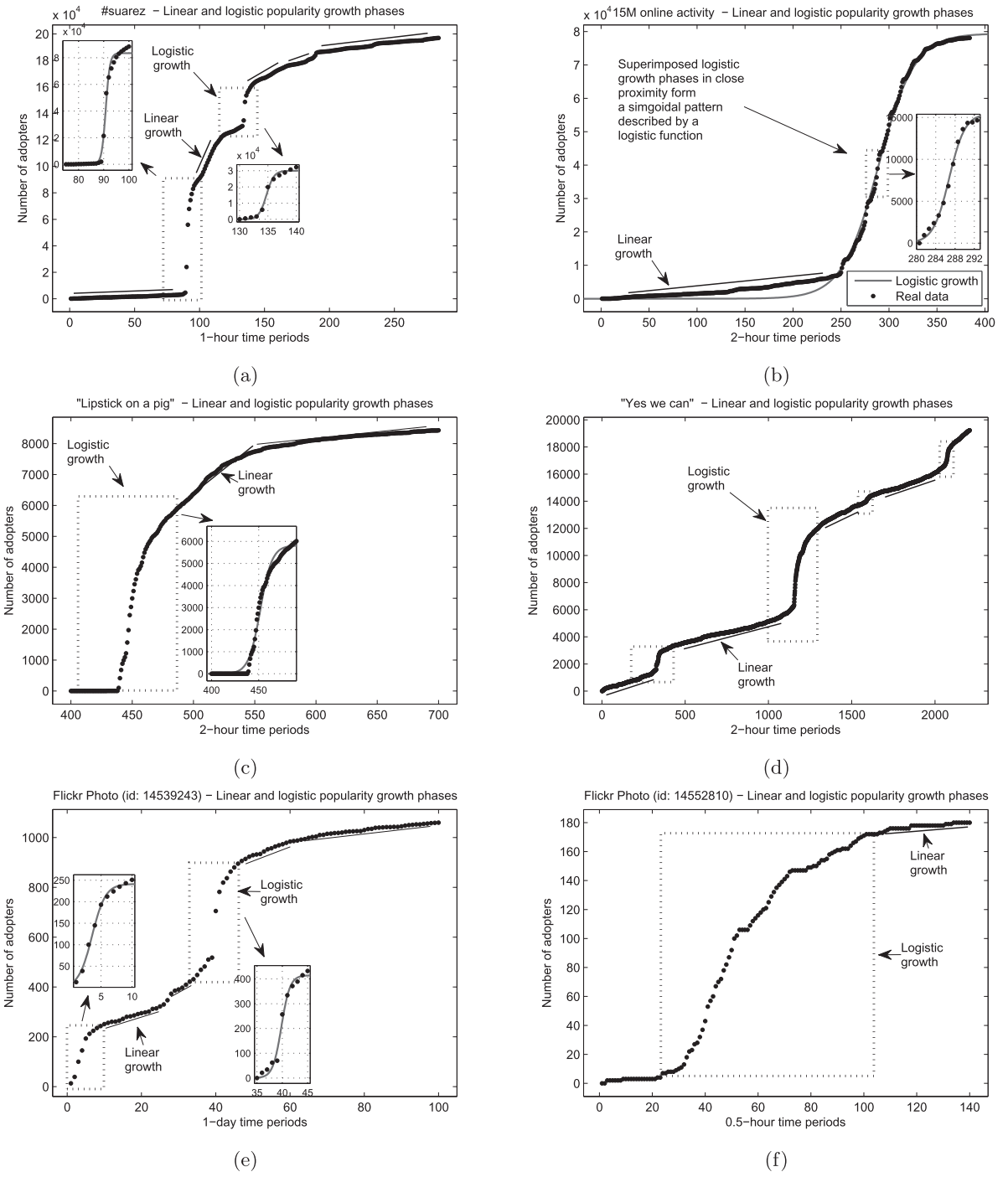

Fig. 1. Popularity evolution of online content through linear and logistic growth phases. (a) #suarez, (b) adopters of the 70 hashtags of the 15 M movement, (c) "Lipstick on a pig" memetracker phrase, (d) "Yes we can" memetracker phrase, (e) - (f) Flickr photos.

2. Related work

The prediction of the popularity of online content is still an open problem. Thus far, many methods have been proposed using different approaches to the formulation of the prediction problem and the techniques employed to provide forecasts. A major research strand mainly focuses on the attributes of the online content for the estimation of its future popularity. In keeping with this research path Ma et al. [28] developed a method for the prediction of the popularity of newly appear-

ing hashtags, based on the classification of the content features of messages, which were used as predictors of popularity trends. Jenders et al. [20] studied the dynamics of viral tweets through the investigation of content and users' characteristics which were employed in the construction of a learning predictive mechanism. Bandari et al. [3] examined the problem of popularity forecasts before the online dissemination of content by using article features as indicators of their future popularity. Asur et al. [2] investigated the factors underlying the formation and persistence of trends in the diffusion of online content. This study suggested that the amplification of content significance among online users constitutes a primary factor in the development of popularity trends. Gupta et al. [18] employed classification and hybrid methodologies along with a set of quality features in predicting the future popularity of events discussed in microblogging social media. Naveed et al. [31] studied the evolution of retweets through a content-based approach, whereby the content characteristics of Twitter messages were used as proxies of their interestingness. Tsagkias et al. [38] proposed a method for the popularity prediction of news articles based on a two-step binary classification task. In the first stage the method predicts the potential of a news article to receive comments. Subsequently, it provides a dichotomous estimation (high-low) of the volume of comments before the online dissemination of an article. Petrovic et al. [32] examined the propagation mechanisms of Twitter messages by focusing on the prediction of retweet patterns through a machine learning approach relying on the social, as well as on the content characteristics of messages. Vasconcelos et al. [41] proposed a featured-based method for the prediction of the popularity of reviews in Foursquare using information about their creators, the place they refer to, and the review content.

Another research trend lays emphasis on the social network topological properties, the users' characteristics and their behavioral pattern. Cui et al. [11] applied this approach to the cascade prediction problem by using properly selected nodes in a social network as sensors of imminent activity bursts. Kupavskii et al. [22] examined the prediction of retweet cascades and proposed a model for estimating the number of retweets over time using the activity flow and the PageRank of retweet graphs. Tsur & Rappoport [39] developed an approach combining content, network, and temporal characteristics as predictors of the diffusion patterns of Twitter hashtags. Zaman et al. [47] used probabilistic collaborating filtering on data pertaining to the identities of the creators of Twitter messages and of their repeaters, with a view to predicting the growth of retweets. Lerman & Hogg [24] used stochastic user behavior models enhanced with additional factors, such as portal aesthetics and initial online activity, in order to forecast the popularity of news on Digg. Bao et al. [4] investigated the role of social network characteristics and particularly the structural diversity of early adopters in order to estimate the popularity of online messages. Weng et al. [43] used diffusion patterns in relation to the topological and community characteristics of a social network as popularity predictors of memes. Cheng et al. [8] proposed a method for the prediction of reshare cascades on Facebook using temporal and network structure indicators as predictors of the sizes of information cascades. Zaman et al. [46] examined the retweet dynamics using a Bayesian approach to the development of a prediction method based on retweet time-series, and the network topological characteristics of retweeters. Zhang et al. [48] used content and context characteristics of Twitter messages, such as the users' behavior and the topological properties of their network position, for the prediction of the messages popularity.

Many studies have examined the popularity prediction problem from a temporal perspective with a view to identifying common patterns allowing the prediction of popularity trends. Crane & Sornette [10] analyzed the propagation patterns of Youtube videos and identified four elementary classes which were modeled through a self-exciting Hawkes process featuring four elementary kernels. Yang & Leskovec [45] studied the temporal patterns of the diffusion of online content and identified common trends in the growth and decline of their popularity, which were classified into six key patterns. Matsubara et al. [29] investigated the rising and falling trends in the popularity of online content and introduced a method for the modeling and prediction of activity patterns. Figueiredo [13] proposed a methodology for the prediction of the popularity of user generated videos relying on the similarities of popularity growth patterns among online videos with common content characteristics. Gursun et al. [19] produced forecasts about the popularity of online videos based on a classification of access patterns according to the activity frequency. Ahmed et al. [1] used a pattern classification approach to the prediction of the popularity of user generated content. By identifying general temporal characteristics of popularity trends, the proposed method predicted the popularity evolution by associating the observed trends with the dynamics of predefined classes. Kong et al. [21] by analysing temporal, content and user characteristics of twitter messages containing specific hashtags, proposed a solution to the problem of the real time prediction of future activity bursts and of their sizes. Li et al. [26] measured and analyzed viewing patterns of online video and proposed a method for the prediction of popularity peaks based on the observed trends and the interestingness of video content. Shen et al. [35] employed a reinforced Poisson process in the prediction of the popularity growth of online content based on the modeling of the temporal pattern of the arrival of new adopters. Building on this approach, Gao et al. [14] modeled the evolution and the prediction of the retweets of a message through an extended reinforced Poisson process, and a time transformation procedure for the elimination of circadian temporal variations in the users' activity. Zhao et al. [49] studied the prediction of the popularity of Twitter messages through a doubly stochastic self-exciting point process. The proposed method modeled the infectiousness of the initial message and the intensity of retweets as functions of time, of the users' connectivity, and of a kernel representing the distribution of the users' reaction times.

Several studies use the initial popularity of online content as predictor of its future evolution. Szabo & Huberman [36] relied on early activity patterns for the estimation of the long-term popularity of online content. Forecasts were generated on the basis of the linear correlation between the logarithms of the early and later popularity of online content. Expanding this model Pinto et al. [33] used a multivariate linear regression method for a more accurate estimation of the early popularity trend of Youtube videos, through multiple sampling of the initial activity using equally sized time intervals. Wu et al.

[44] used the access history of online content as input to a reservoir computing technique providing estimates about its future popularity. Lee et al. [23] approached the popularity prediction of online content through survival analysis based on publicly available information about the initial popularity trend. A useful resource for further information on models for the popularity prediction of online content is the survey by Tatar et al. [37].

3. Modeling the popularity growth of online content

In this section we analyze the empirical findings with a view to identifying universal patterns explaining the phenomenological observations, and enabling a general modeling approach to the popularity growth of online content.

3.1. Modeling framework

Both the linear and the sigmoidal popularity growth periods have a phenomenological importance concerning the understanding of the underlying adoption dynamics of online content. During the linear growth periods, the evolution of the adoption rate (new adopters per time period) undergoes a steady state phase, due to the statistical equilibrium of the adoption dynamics. In a state of statistical equilibrium the properties of online content, as well as the characteristics and habits of potential adopters do not change over time, which coupled with the absence of interventions from the external environment lead to a stationary adoption pattern. In such a phase, new adoptions occur through minor information cascades stemming from the independent online activity - such as content searching and sharing of links - of a population of users with homogeneous adoption thresholds reflecting highly common preferences, interests, needs, habits and demographic characteristics. Such properties can be considered constant within the evolution timescale of the online content diffusion process, thus leading to a stable adoption probability limiting the penetration of online content to a homogeneous subpopulation of low threshold users. It is worth mentioning that a process might be stationary when examined in one timescale and non-stationary in another. For instance the adoption rate of a Twitter hashtag during a day may present variations due to the users' biological clock and their daily schedule. When examining the adoption process using timescales of minutes or hours the activity pattern will be non-stationary. The shorter the timescale, the more changeable the adoption rate will be. However, the same phenomenon in a timescale of days will be stationary, provided a hashtag is persistent and the daily adoption pattern is repeated, thus leading to small variations in the number of new adoptions per day. Similarly, stationarity might be detected in larger timescales, such as weeks or months, as long as the weekly or the monthly pattern is repeated. When the online content under examination has an international audience, then the hourly variations in the adoption rate that would appear if the content was of local interest are smoothed, due to the different time zones of the users' locations. In such a case, the adoption rate might be stationary even in smaller timescales.

In stationary phases the adoption rate presents minor fluctuations around a mean value, thus making the content popularity grow in a linear way. The linear trend can be captured by a 1^{st}-degree polynomial model of the form $N(t) = P \cdot t + N(t_0)$, whose independent variable is the time t, and the dependent variable is the total number of adopters $N(t)$ at time t. The single parameter P of this model represents the stationary adoption rate, that is, the number of adopters per time period (e.g. bins of minutes, hours, days, weeks, etc.). A stationary popularity growth trend is always present, since it captures the online content diffusion dynamics among the low threshold users. An intervention event causes perturbations to the adoption rate, as users who normally would not adopt the propagating online content, that is, higher threshold users, become adopters due to the excitatory effect of an external event [27]. Actually, the reason underlying the change in the adoption rate when an intervention event takes place, is the increased susceptibility of heterogeneous users who are brought closer to their adoption threshold, and not a change in the contagiousness of the online content, insofar as its characteristics remain the same. This means that the new adoptions triggered by an intervention event are analogous to opportunistic infections caused by a virus taking advantage of the occasional increased vulnerability of a subpopulation of individuals. For instance, the characteristics of a Youtube video of a famous singer's song are the same before, during, and after a popularity burst caused by an important event, such as a concert, or whatever increases the singer's publicity. Certainly, before, during, and after a burst, new views of the video under consideration come from users who are strongly identified with the artist, however during the burst other users with different characteristics also watch the video due to the occasional excitation generated by the sudden public interest. While the former category of users continue being susceptible to this video, thus functioning as a signal propagation medium through which the stimuli generated by the video travels within a social network, users form the latter category will not be a pool of new adopters, as their susceptibility was transient and exogenously driven.

In essence, an intervention event dynamically changes the number of potential adopters, thus temporarily increasing the carrying capacity of the environment within which the popularity of online content can grow. Considering that individuals are characterized by heterogeneous adoption thresholds, assumed to follow a unimodal, or even normal distribution according to the diffusion of innovations theory by Rogers [34], then the adoption process initiated by an intervention event proceeds as follows: In the beginning those who become adopters belong to the low-threshold subpopulation. The susceptibility of these users is close to that of the users sustaining the stationary trend, but at the same time higher than the level required for being adopters during the stationary phase. Otherwise, the stationary adoption rate would be higher; the width of the distribution of higher threshold users would be narrower; and the size of the adoption bursts due to external events smaller. As the adoption process unfolds, new users with thresholds closer to the mode of the threshold distribution are

progressively recruited. While the adoption cascade approaches the mode of the users' threshold distribution by sweeping its left part, its size is proportional to the increasing number of potential adopters who are concentrated in this threshold distribution area. This adoption phase forms the lower part of the sigmoidal popularity growth. Subsequently, as the adoption cascade moves away from the mode of the threshold distribution – which corresponds to the inflection point of the sigmoidal adoption pattern – the number of potential adopters progressively decreases and the popularity growth slows down. A unimodal distribution of thresholds gives rise to a sigmoidal popularity evolution pattern, where the total number of adopters $N(t)$ at time t is obtained by the logistic function described by the following mathematical formula:

$$N(t) = \frac{C}{1 + e^{(-a(t - t_{infl}))}} \tag{1}$$

Eq. (1) contains three parameters: First, the saturation level of the sigmoidal curve, denoted by C. This value represents the carrying capacity of the environment, that is, the number of individuals who have become susceptible to the propagating online content. The value of the parameter C depends on how interesting is the external intervention and its intensity. The second parameter of Eq. (1) represents the steepness of the sigmoidal curve, denoted by a. We can conceptualize the gradient of this curve as an indicator of the level of tuning of the adopters to the content they have been exposed to. In other words, the slope of the sigmoidal curve relates to the variance of the distribution of the users' adoption thresholds. The lower the variance, the higher the slope will be, and vice versa. The third parameter of Eq. (1), denoted by t_{infl}, is the inflection point of the sigmoidal curve, that is, the time point at which the popularity growth starts slowing down after the initial phase of accelerating increase.

Popularity growth of fleeting, non-popular online content. The popularity of online content regardless of the level and duration of its interestingness, evolves through linear and non-linear phases. Nevertheless, in the case of fleeting online content the stationary adoption phases are significantly shorter in comparison with analogous phases in the propagation of enduring online content, since the interestingness of fleeting content is volatile and short-lived. Inasmuch as a fleeting piece of online content is also non-popular, then there will be additional quantitative differences in comparison with the adoption pattern of popular content. In particular the size of the non-linear adoption cascades caused by intervention events, and the stationary adoption rate of the linear popularity growth phases, will be significantly smaller in relation to the values of these indicators in the case of popular and enduring online content. These quantitative differences are due to the small number of users who are interested in a non-popular piece of online content.

The proposed modeling framework is quite generic and flexible enough to capture the adoption dynamics of both enduring and fleeting, popular and non-popular online content through a universal formulation. More specifically, the popularity growth of fleeting online content after the termination of the adoption process can be modeled in a way analogous to the popularity growth of enduring online content. However, in the former case the last linear popularity growth phase has a zero slope, since no other adoptions take place after a certain time point. In experiment 7 we demonstrate the effectiveness of this approach through the prediction of the popularity growth of a fleeting, non-popular hashtag. In experiment 3 we provide another example of predicting the adoption of a non-popular but enduring piece of online content.

3.2. Statistical explanation of the popularity growth pattern

As already discussed, the logistic function accurately describes the non-linear popularity growth of online content, while a 1st-degree polynomial term captures the dynamics of a linear growth phase. This means that a function comprising a constant term along with dynamic terms, featuring the mathematical structure of the probability density function of the logistic distribution, should fit the adoption rate of a piece of online content containing multiple bursts superimposed on a stable adoption trend. This function takes the form of Eq. (2), where the term p accounts for a stationary adoption rate, the scale factor b_m regulates the height of the mth adoption burst culminating at the time point t_m, and the parameter s_m adjusts the width of the adoption burst.

$$n(t) = p + \sum_m b_m \frac{e^{-\frac{t - t_m}{s_m}}}{s_m (1 + e^{-\frac{t - t_m}{s_m}})^2} \tag{2}$$

The fitting of the function described by Eq. (2) to the adoption rate of the memetracker phrase "Yes we can" is highly precise as shown in Table 1. In this case the adoption rate refers to the number of new urls per time period where this phrase appeared at least one time. The matching between the real data and the values generated by Eq. (2) – both normalized to a reference scale of 100 – is visually illustrated in Fig. 2.

3.3. A model of the popularity growth of online content

Suppose that the popularity evolution of a piece of online content consists of k linear and f sigmoidal growth periods. Let P_k denote the adoption rate of the k^{th} stationary phase, and t_k^s, t_k^e denote its start and end time points correspondingly. Also, let C_f, a_f and t_{infl}^f denote the total size of the f^{th} sigmoidal growth phase, its slope, and its inflection point respectively. Considering that at any time point t there is an ongoing stationary phase (the k^{th}), and in view of the fact that the total

Table 1
Fitting of a model based on the logistic probability distribution function to the adoption rate time-series of the "Yes we can" memetracker phrase.

Fitting of a model based on the probability density function of the logistic distribution to the real adoption rate time-series
Adoption rate model : $n(t) = p + b1 * ((exp(-(t - t_{m1})/s1))/(s1 * (1 + exp(-(t - t_{m1})/s1))^2)) +$ $b2 * ((exp(-(t - t_{m2})/s2))/(s2 * (1 + exp(-(t - t_{m2})/s2))^2)) + b3 * ((exp(-(t - t_{m3})/s3))/(s3 * (1 + exp(-(t - t_{m3})/s3))^2)) +$ $b4 * ((exp(-(t - t_{m4})/s4))/(s4 * (1 + exp(-(t - t_{m4})/s4))^2)) + b5 * ((exp(-(t - t_{m5})/s5))/(s5 * (1 + exp(-(t - t_{m5})/s5))^2))$
Coefficients: $p = 2.162, b1 = 105.71, b2 = 312.34, b3 = 52.86, b4 = 38.44, b5 = 96.1, t_{m1} = 28, t_{m2} = 97, t_{m3} = 102, t_{m4} = 133, t_{m5} = 173, s1 = 0.75, s2 = 0.75, s3 = 0.55, s4 = 0.75, s5 = 0.75$
Goodness of fit: R-square = 0.9978, Adjusted R-square = 0.9978, RMSE = 0.4301

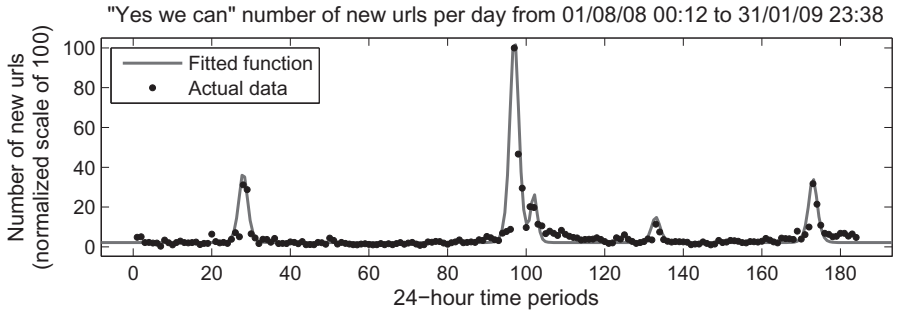

Fig. 2. Fitting of a function based on the logistic distribution formulation to the adoption rate of the "Yes we can" memetracker phrase.

Table 2
Description of the symbols of the proposed model described by Eq. (3).

Symbol	Description
$N(t)$	Cumulative number of adopters at time t
$N(t_0)$	Cumulative number of adopters at time t_0
k	Number of linear growth periods
f	Number of non-linear growth periods
P_j	Adoption rate of the j^{th} linear growth period
P_k	Adoption rate of the current (k^{th}) linear growth period
t_j^s	Start time point of the j^{th} linear growth period
t_j^e	End time point of the j^{th} linear growth period
t_k^s	Start time point of the current (k^{th}) linear growth period
C_i	Total number of adopters during the i^{th} non-linear growth period
a_i	Slope of the i^{th} sigmoidal adoption pattern
t_{infl}^i	Inflection point of the i^{th} sigmoidal adoption pattern

number of adopters at time t_0 was $N(t_0)$, then the total number of adopters $N(t)$ at time t can be mathematically expressed through Eq. (3), whose symbols are summarized in Table 2.

$$N(t) = N(t_0) + \sum_{j=1}^{k-1} P_j \cdot (t_j^s - t_j^e) + P_k \cdot (t - t_k^s) + \sum_{i=1}^{f} \frac{C_i}{1 + e^{(-a_i(t - t_{infl}^i))}} \tag{3}$$

Eq. (3) describes the popularity evolution of an enduring piece of online content through multiple linear and non-linear growth periods. The same equation can be used in the description of the popularity growth of fleeting online content. In such a case, the linear term $P_k \cdot (t - t_k^s)$ corresponds to a stationary adoption phase of zero growth rate ($P_k = 0$). As a result, we maintain the general equation-based model of the online content popularity growth, meanwhile accounting for the fact that the adoption process might end during the observation time window.

Inasmuch as the stationary adoption rate remains constant during the examined period, then the model described by Eq. (3) becomes simpler and takes the form of Eq. (4). The r.h.s term of Eq. (4) consists of the total number of adopters $N(t_0)$ at t_0, a 1st-degree polynomial term describing the linear popularity growth due to a stable stationary trend; and a sum of logistic terms accounting for the non-linear popularity growth phases due to intervention events. In the case of a fleeting piece of online content featuring a single stationary trend ending within the observation time window, the popularity growth is expressed through Eq. (3) with $k = 2$. The first stationary trend accounts for the non-zero linear popularity growth during the statistical equilibrium of the adoption dynamics. The second stationary trend pertains to the zero popularity growth after the termination of the adoption process. Experiment 7 provides an example of this case.

$$N(t) = N(t_0) + P \cdot t + \sum_{i=1}^{f} \frac{C_i}{1 + e^{(-a_i(t - t_{infl}^i))}} \tag{4}$$

4. A method for predicting the popularity of online content

Having constructed a model describing the popularity evolution of online content, we shift our focus to the estimation of its parameter values, with a view to generating forecasts through forward extrapolation.

Prediction problem formulation. Using as input information the adoption rate of a piece of online content c, that is, the number of new adopters $n(t_i)$ per time period $t_i (i = 0, 1, 2, \ldots)$ of configurable temporal resolution (e.g. bins of minutes, hours, days), is it possible to predict the total number of adopters $N(t_p)$ at a time point $t_p > t_i$?

Explanation of the popularity prediction method. We consider that a user, or another entity e.g. a url, becomes an adopter of the online content c at the time of its first use, and then permanently stays in the "adopter" state. Consequently, the cumulative number of adopters monotonically increases through a birth process ($dN(t)/dt = births$) until the last adoption event.

During periods of linear popularity growth the birth rate is constant, however this rate dynamically changes during non-linear growth phases. In such phases the birth rate is the first derivative of the logistic function (Eq. (1)), that is, $dN(t)/dt = f(N(t)) = a \cdot N(t) \cdot (1 - N(t)/C)$. While the first derivative of the birth rate (the second derivative of the logistic function) is positive, that is, $\frac{df(N(t))}{dt} = \frac{d^2N(t)}{dt^2} > 0$, the adoption process accelerates. When it becomes negative ($\frac{d^2N(t)}{dt^2} < 0$) the adoption process decelerates. The time point at which the quantity $acc = \frac{d^2N(t)}{dt^2}$ changes trend from positive to negative, corresponds to the inflection point of the sigmoidal popularity growth curve.

By monitoring the acceleration acc of the adoption rate of the online content c, and by evaluating whether it fluctuates within a range $[-\Delta, \Delta]$, we can identify the popularity growth mode (linear or non-linear), and accordingly estimate the model parameters. In order to absorb the minor fluctuations of acc around zero during stationary adoption periods, we define a range $[-\Delta, \Delta]$ within which we consider the fluctuations insignificant. The proper tuning of Δ can be achieved through experimentation aiming at assessing the accuracy of the predictions against real data for various values of Δ. When the value of acc crosses Δ while increasing, the adoption process becomes non-stationary, meaning that an adoption cascade larger than the usual ones is in progress. In such a phase, the popularity growth of the online content c follows a non-linear, sigmoidal pattern described by the logistic function (Eq. (1)). For the estimation of the parameters of this function we need the cumulative number of adopters from the penultimate time point t_s at which $acc \in [-\Delta, \Delta]$ to the time point $t^{first}_{acc<0}$ at which acc becomes negative for the first time after crossing the threshold value Δ while increasing at time $t^{first}_{acc>\Delta}$. Between the time point $t^{last}_{acc>0}$ at which acc was positive for the last time after $t^{first}_{acc>\Delta}$, and the time point $t^{first}_{acc<0}$ lies the time point t_{infl} corresponding to the inflection of the sigmoidal popularity growth pattern. Through experimentation we found two approximations of t_{infl}. One estimate is the time point at which the graph of $acc = \frac{d^2N(t)}{dt^2}$ crosses the zero axis as it changes trend from positive to negative. The other estimate of t_{infl} coincides with the time point $t^{last}_{acc>0}$.

5. Experimental results

In this section we test the effectiveness of the proposed method in predicting the popularity of online content and assess its performance against state-of-the-art methods. We demonstrate that the proposed method is substantially more accurate, significantly less demanding on information, and answers more prediction questions than the benchmark methods.

5.1. Benchmark models

Szabo & Huberman model (S-H). According to this model [36] the popularity of online content evolves as a stochastic process driven by random events adding a small arbitrary amount to the logarithm of the cumulative popularity. The model is based on the significant linear correlation between the logarithmic transformations of the initial and final popularity of online content.

Multivariate Linear Model (ML). Pinto et al. [33] built on the S-H model [36] and increased the prediction accuracy by assigning different weights to the popularity samples obtained during the early activity period. Forecasts about the popularity of a video are produced through a multivariate linear regression using the estimated early popularity parameters, and regression coefficients between the initial and later popularity, derived from the training of the model on empirical data.

MRBF model. Pinto et al. [33] proposed the MRBF model – an enhanced, more powerful version of the ML model. The MRBF model takes into consideration the similarity of the early popularity of a specific online video with the popularity trends of videos existing in a training dataset. The similarity is measured by a Radial Basis Function (RBF).

Table 3
Qualitative comparison of the proposed prediction method with the benchmarks.

Requirements	Proposed model	SEISMIC	Kong et al.	S-H	ML	MRBF
Training	No training	Training requiring the users' reshare activity history, node degrees, estimation of the waiting time distribution.	Training using 2/3 of data containing meme, user, content, network, hashtag, time-series, and prototype features.	It requires training for the calculation of the regression coefficients between t_i (target time) and t_r (reference time).	Training is required. Because of the different weights assigned to the samples of the early popularity, the process is more complicated in relation to that of the S-H model.	It requires a training dataset with early popularity patterns for measuring the similarity of the initial popularity trend with the reference patterns.
Real time	Yes	Yes	Yes	No	No	No
Microlevel - Macrolevel	Macrolevel - only aggregate popularity data is required.	Microlevel - it models the dynamics at a user level.	Microlevel	Macrolevel - only aggregate popularity data is required.	Macrolevel - only aggregate popularity data is required.	Macrolevel - only aggregate popularity data is required.
Modeling and prediction of popularity bursts	Yes	No	Yes - logarithm of burst size	No	No	No
Content feature analysis	No	No	Yes	No	No	Yes
Knowledge of network structure	No	Limited (node degrees, activity per node)	Yes	No	No	No

SEISMIC model. The model by Zao et al. [49] deals with the real time prediction of the final number of retweets of a Twitter message. The model is based on a self-exciting point process producing popularity forecasts using the early reshare activity of a particular message. The model addresses the popularity growth dynamics at a user level, thus requiring the estimation of a memory kernel accounting for the distribution of the delay times between the arrival of a message to a user and its retransmission. Also, the model requires the estimation of the infectiousness of a Twitter message, its reshare time points and the out-degree of the resharers.

Kong et al.. The method by Kong et al. [21] aims at predicting whether a hashtag will burst in the future, and if it bursts how large its popularity will be. Predictions about the burst sizes are provided in terms of the natural logarithms of the actual sizes. Forecasts about whether a hashtag will burst or not, are generated through a binary classification task based on a weighted support vector machine approach. The supervised learning process examines user, content, network, hashtag and time-series features, as well as historical information to detect similarities with the hashtags which are monitored.

5.2. Qualitative comparison of the proposed method with the benchmarks

In Table 3 we qualitatively compare the proposed method with the benchmarks, with a view to highlighting the differences in terms of information requirements, types of forecasts, training period and technical characteristics. The comparison shows that the proposed method is substantially simpler, in that it does not require training, nor it needs micro-level information, such as out-degree of nodes, distribution of waiting times, or content features analysis. In Table 4 we summarize the prediction questions addressed by the proposed method and reveal its increased predictive capacity in comparison with the benchmarks.

5.3. Performance assessment metrics

In the comparison with the *S-H*, *ML*, *MRBF* models we use the *Relative Square Error (RSE)* metric, so the performance of the proposed method is evaluated against the results reported in Pinto et al. [33]. Let $N(t_i)$ denote the cumulative number of adopters of a specific online content at time t_i $(N(t_0) = 0)$, and let $\widehat{N(t_i)}$ denote the prediction about the total number of adopters at time t_i, the *RSE* is calculated by the formula $RSE(t_i) = (\widehat{N(t_i)}/N(t_i) - 1)^2$. In the comparison with the SEISMIC

Table 4
Prediction questions answered by the proposed and the benchmark methods.

Prediction question	Proposed model	SEISMIC	S-H	ML	MRBF	Kong et al.
Q1: Can we predict the popularity evolution of online content?	Yes in linear and non-linear growth periods	Yes in linear growth periods	Yes in linear growth periods	Yes in linear growth periods	Yes in linear growth periods	Yes in linear and non-linear growth periods
Q2: Given a time point at which we expect an adoption peak how early, and how accurately can we forecast: a) the final size of the adoption cascade, b) the evolution of the adoption pattern?	Yes	No	No	No	No	No
Q3: Can we predict the popularity of various types of online content? (universality)	Yes	No	No	No	No	No
Q4: Can we determine if a cascade is in the bursting or the decaying phase?	Yes	No	No	No	No	Yes

model we use the *Absolute Percentage Error (APE)* as in Zhao et al. [49]. Using the foregoing notation the *APE* is given by the formula $APE(t_i) = |(\widehat{N(t_i)} - N(t_i))/N(t_i)|$. Another standard statistical error metric that we use for measuring the goodness-of-fit of the predictive model to the actual data, is the *Root Mean Square Error (RMSE)*. Let $N^{est}(t_i)$ denote the values estimated by the model, and let $N^{obs}(t_i)$ denote the values of the observations, then for k pairs of actual and estimated values, the *RMSE* is calculated by the formula $RMSE = \sqrt{(1/k)\sum_{i=1}^{k}(N^{obs}(t_i) - N^{est}(t_i))^2}$.

5.4. Description of datasets and data analysis

Twitter messages: We analyzed sequences of Twitter messages containing specific hashtags from three different datasets. The first one was compiled by Weng et al. [43], who used it in their study on the virality prediction of internet memes. This publically available dataset consists of approximately 1.35 million hashtag sequences created by almost 1.5 million users from March 24, 2012 to April 25, 2012. The messages of this dataset constitute a 10% sample of the public tweets created during the aforementioned period. Each item of a hashtag sequence contains two information elements, namely, the message timestamp and the anonymized id of its creator. The second Twitter dataset relates to the 15 M protests in Spain. This dataset was used in the study of the online protest dynamics of the 15 M movement in Spain by González-Bailón et al. [16]. The dataset comprises 529,393 Twitter messages created by 78,080 users from 00:03:26 April 25, 2011 to 23:41:23 May 26, 2011. The messages of the dataset contain hashtags from a list of 70 ones referring to the 15 M protests. Each message consist of the user id (anonymized), the creation timestamp, the hashtags used in the message, and the ids of the users who were mentioned in it (if any). The third Twitter dataset consists of sequences of public Twitter messages containing the hashtag #Suarez. This dataset contains 260,878 timestamped messages created by 178,750 users in the period from 00:00:06 June 21, 2014 to 13:00:06 June 28, 2014. During this period Luis Suarez' unsporting behavior in the match between Italy and Uruguay in the World Cup of 2014 caused activity bursts, as users commented on this incident. We downloaded these messages through the Scraperwiki[1] platform. Subsequently the messages were processed in google-refine[2] - a data cleansing tool - in order to remove unnecessary information, while maintaining the timestamp and the user id, which was anonymized.

Memetracker phrases: In addition to sequences of Twitter messages we analyzed time-series referring to the appearances of specific phrases in online articles and blogs. This data was obtained from Memetracker [25] – a publically available dataset. Memetracker[3] consists of phrase clusters extracted from online articles and blog posts. From the Memetracker dataset we processed phrases from the phrase cluster data. For each phrase in a cluster, the dataset contains information about the time of its first appearance in a url, the type of the url (blog or mainstream media), the url address, and the number of appearances in this url. In total there exist around 8.5 million entries of this type in the phrase cluster data of the Memetracker dataset. The processed data covers the period from August 1st, 2008 to January 31st, 2009.

Flickr dataset: To provide further evidence of the universality of the proposed approach, we analyzed data relevant to the evolution of the popularity of photos uploaded on Flickr. The dataset that we used was part of the dataset analyzed in a study on the online information propagation by Cha et al. [7]. In particular, from the entire dataset of the study we analyzed

[1] http://scraperwiki.com.
[2] https://code.google.com/p/google-refine/.
[3] http://www.memetracker.org/data.html.

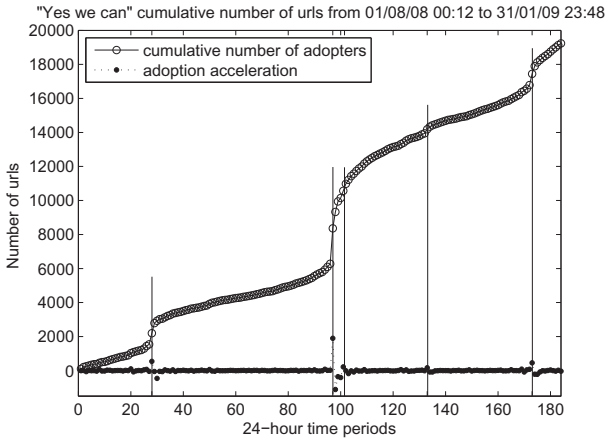

Fig. 3. "Yes we can": Popularity evolution and acceleration of the adoption rate per 24-hour periods.

information from the part containing all the favorite markings on the Flickr photos of the dataset. The dataset comprises 11,267,320 photos which gathered 34,734,221 favorite markings. Each row in the list of favorite markings contains the id of the user who marked a photo as favorite, the id of the marked photo, and the timestamp of the favorite marking. All these information fields were anonymized. The data was collected during two distinct time periods, with the first one extending from November 2nd to December 3rd, 2006; and the second one from February 3rd to May 18th, 2007.

Data processing: From the aforementioned datasets we extracted three types of time-series, namely, the evolution of the cumulative number of adopters of a piece of online content (i.e. Twitter hashtag, Memetracker phrase, Flickr photo), the evolution of the adoption rate (number of new adopters per time period), and the first derivative of the adoption rate (adoption acceleration). In the data processing we used parametrically defined temporal resolutions (time bins of minutes, hours etc.), and chronological periods (start - end points of the time-series).

5.5. Experiments

By applying the proposed method to the prediction of the popularity of various types of online content, we demonstrate its effectiveness and the significant improvement on the performance of the benchmark models. Information about the precision of the predictions of the S-H, ML, MRBF, and SEISMIC models was obtained from the performance reports (Tables and Figures) provided in the studies by Pinto et al. [33], and Zhao et al. [49]. Apart from measuring the accuracy of the predictions, we also evaluated the goodness-of-fit of the proposed model to real popularity growth patterns. Since the prediction method relies on that model for generating forecasts, the excellent fit of the model to empirical popularity patterns and the normality of the residuals, confirm its suitability for generating reliable popularity predictions.

5.5.1. Experiment 1: Prediction of the popularity of the "yes we can" memetracker phrase

The first experiment aims to predict the popularity of the "Yes we can" memetracker phrase. The values of the popularity time-series pertain to the evolution of the total number of urls, where the phrase "Yes-we-can" appeared for at least one time. The analyzed data covers the period from 01/08/08 00:12 to 31/01/09 23:48. In total there were 19,233 urls which adopted this phrase. The popularity growth per 24-hour periods and the corresponding acceleration of the adoption rate are depicted in Fig. 3.

Suppose we start the prediction process at t_0. To produce a popularity forecast we need to understand whether the phenomenon evolves in a stationary or non-stationary way. To this end, we observe the acceleration rate during the first five days. As it can be seen from the corresponding part of Fig. 3, the acceleration presents small fluctuations around zero, thus implying a stationary adoption pattern which can be captured by a 1st-degree polynomial model of the type $N(t) = P_1 \cdot t + N(t_0)$. The parameter P_1 accounts for the stationary adoption rate. Using the cumulative number of adopters (urls) of the first five days (length of reference time period $t_r = 5$) we can estimate the 1st-degree polynomial model and make forecasts about the popularity of the phrase. As long as the adoption acceleration remains within the stationarity bounds, the predictions are valid and the model can continue producing forecasts through a linear extrapolation of the current trend. Fig. 4a depicts the forecasts generated by a 1^{st}-degree polynomial model overlaid on the real data. The forecasts extend until the target time point $t_i = 27$, that is, the 27th day of the observed phenomenon. The APE and RSE values of the predicted popularity are illustrated in Fig. 5a and b respectively. The APE and RSE values of the popularity prediction at the target time point t_{27} are equal to 0.123 and 0.015 respectively. Both values are better than the performance of the SEISMIC, S-H, and ML models in linear prediction tasks of analogous t_r/t_i ratio, as illustrated in Table 5.

At time t_{28} we observe a rapid increase in the acceleration rate exceeding the upper bound of the stationarity margin Δ (in this case Δ=100). This unusual acceleration of the adoption rate corresponds to a non-linear popularity growth phase,

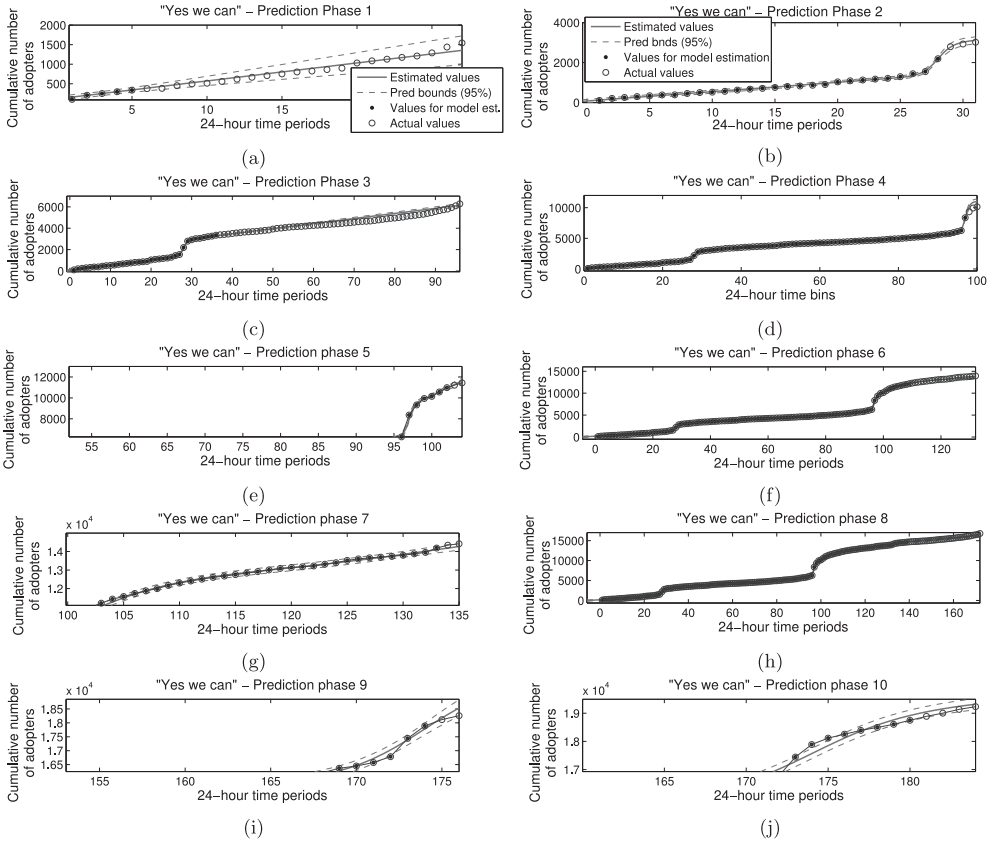

Fig. 4. "Yes-we-can": (a)–(j) Prediction phases 1-10. Real values of the cumulative number of adopters (urls) overlaid on the estimated values. The legends of (a) and (b) apply to all the other subfigures.

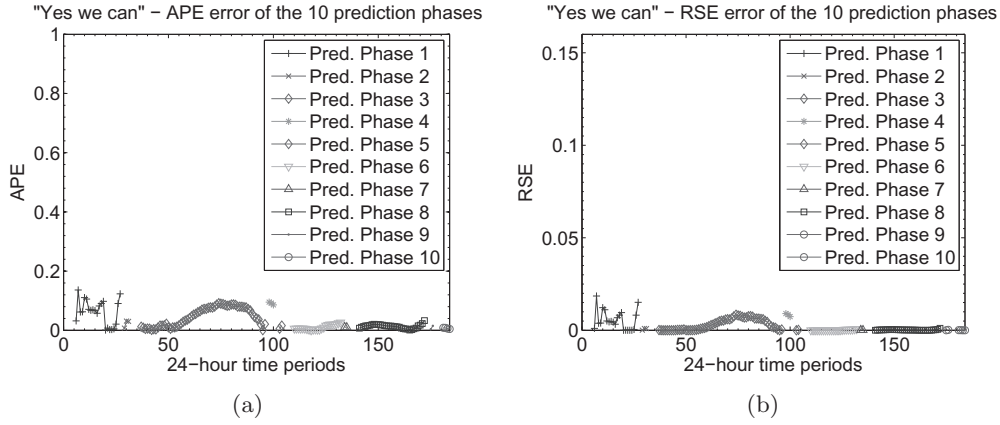

Fig. 5. "Yes we can": (a) APE values of the ten prediction phases, (b) RSE values of the ten prediction phases.

whose evolution was also predicted by the proposed method. A non-linear popularity growth phase is modeled by a logistic function with an intercept equal to the total number of adopters at time t_{27}, that is, the outcome of a linear popularity growth phase recruiting adopters from the most susceptible subpopulation. Entities (urls in this case) from this subpopulation continue adopting the "Yes we can" phrase with the same rate during the ongoing non-linear adoption phase. This means that the cumulative adoption growth incorporates two trends, namely, the stationary trend recruiting adopters from the low threshold population, and the non-linear trend recruiting opportunistic adopters from a higher threshold population. As a result the model in the 2nd prediction phase takes the following form:

$$N(t) = N(t_0) + P_1 \cdot t + \frac{C_1}{1 + e^{-a_1(t - t^1_{infl})}} \tag{5}$$

Table 5
"Yes we can": Accuracy of the predictions at target time points and comparison with the benchmark methods. The numbers in parentheses indicate the performance of the benchmarks in linear prediction tasks. The performance of the benchmark models was estimated according to the (t_r/t_i) ratio of the prediction task, and the error values reported in Zhao et al. [49], and Pinto et al. [33].

Prediction phase #, type, t_r/t_i	Proposed model (APE & RSE)	SEISMIC model (APE)	S-H model (RSE)	ML model (RSE)
1, linear, 0.185	0.123, 0.015	∼ 0.15	∼ 0.24	0.20
2, non-linear, 0.25	0.03, 0.000916	N/A (∼ 0.125, linear)	N/A (∼ 0.25, linear)	N/A (< 0.25, > 0.2, linear)
3, linear, 0.077	0.021, 0.000441	∼ 0.185	∼ 0.4	∼ 0.4
4, non-linear, 0.25	0.085, 0.0072	N/A (∼ 0.125, linear)	N/A (∼ 0.25, linear)	N/A (< 0.25, > 0.2, linear)
5, non-linear, 0.5	0.0157, 0.000246	N/A (∼ 0.1, linear)	N/A (∼ 0.12, linear)	N/A (0.1, linear)
6, linear, 0.217	0.0266, 0.000709	∼ 0.125	∼ 0.25	∼ 0.2
7, non-linear, 0.33	0.0102, 0.000246	N/A (∼ 0.12, linear)	N/A (∼ 0.2, linear)	N/A (∼ 0.15, linear)
8, linear, 0.1351	0.0327, 0.00107	∼ 0.15	∼ 0.4	∼ 0.35
9, non-linear, 0.5	0.0152, 0.000232	N/A (∼ 0.1, linear)	N/A (∼ 0.12, linear)	N/A (∼ 0.1, linear)
10, linear, 0.5	0.0045, 0.0000204	∼ 0.1	∼ 0.12	∼ 0.1

Table 6
"Yes we can": Goodness-of-fit of the model to the data (normalized to a scale of 100).

"Yes we can": Fitting of the model to the data
Model: $N(t) = N(t_0) + P1 * t + (c1/(1 + exp(-a1 * (t - 28)))) + (c2/(1 + exp(-a2 * (t - 97)))) + (c3/(1 + exp(-a3 * (t - 102))))$ $+(c4/(1 + exp(-a4 * (t - 133)))) + (c5/(1 + exp(-a5 * (t - 174))))$
Coefficients (95% confidence bounds): a1 = 0.9283 (0.7929, 1.064), a2 = 1.967 (1.722, 2.211), a3 = 0.233 (0.2204, 0.2456), a4 = 0.1817 (0.1467, 0.2167), a5 = 0.3272 (0.3007, 0.3537) c1 = 9.522 (9.242, 9.802), c2 = 17.54 (17.03, 18.05), c3 = 17.11 (16.38, 17.85), c4 = 5.21 (4.769, 5.652), c5 = 14.93 (14.41, 15.45), $N(t_0)$ = 0.9997 (0.8684, 1.131), P1 = 0.191 (0.1862, 0.1959)
Goodness of fit: R-square: 0.9999, Adjusted R-square: 0.9999, RMSE in a normalized scale of 100: 0.3071

In Eq. (5), $N(t_0)$ represents the intercept of the model at t_0; P_1 denotes the stationary trend; C_1 denotes the size of the 1st non-linear adoption cascade; and a_1 represents the slope of the sigmoidal growth. We denote by t^1_{infl} the inflection point of the 1st sigmoidal popularity pattern. In this case the inflection point is at t_{28} as shown in Fig. 3. To estimate the size of the non-linear adoption cascade, we make a forecast for the next three time points after t_{28}. The actual values along with the predicted ones are depicted in Fig. 4b. We should point out that the prediction of the size of the cascade was made after adding only one data point for the estimation of the model parameters, that is, the data point corresponding to the inflection of the sigmoidal curve (t_{28}). The APE and RSE values of the 2nd prediction phase are illustrated in Fig. 5a and b respectively. The APE and RSE values of the prediction at the target time point t_{31} are equal to 0.03 and 0.000916 respectively. We mention that the SEISMIC, S-H and ML models do not produce forecasts about the size of information cascades during their bursting phase. It is noticeable that the prediction performance of the proposed method in non-linear phases is better than the performance of the benchmark models in linear prediction tasks of analogous t_r/t_i ratio, as shown in parentheses in Table 5.

The first non-linear popularity growth phase is followed by a linear one. As discussed earlier, the stationary propagation trend is generated by the most susceptible subpopulation. To the extent that the stationary trend remains the same, the existing model (Eq. (5)) can still predict the popularity growth. Using only five new data points for the re-estimation of the stationary trend (small adjustment), we generated predictions up to the target time point t_{96}. Figure 4c depicts the actual popularity values overlaid on the estimated ones. Despite the long prediction horizon there is a close fit of the predicted values to the real ones. In the 3rd prediction phase we used five new data points (t_{32}, ..., t_{36}) and estimated the evolution of the popularity growth from t_{37} to t_{96}, which means that we had a prediction task of a t_r/t_i ratio equal to 0.077. The prediction accuracy of the proposed model in terms of APE and RSE is depicted in Fig. 5a and 5 b. The accuracy of the popularity predictions at the target time points is higher than that of the benchmarks, as illustrated in Table 5.

The prediction of the popularity of the memetracker phrase "Yes we can" proceeded by adding a non-linear growth term each time an intervention event caused the adoption process to evolve in a non-stationary way. In this example we had three more non-linear growth phases after the one already described. The results of the remaining prediction phases are depicted in Fig. 4 (from (d) to (j)), Fig. 5a, and b, and Table 5. The final model gives an excellent goodness-of-fit to the entire popularity growth pattern as provided in Table 6. The R-square and the Adjusted R-square are equal to 0.9999, and the RMSE was estimated at 0.3071 after normalizing the data and the predictions to a reference scale of 100. The fitting of the final model to the real popularity pattern is visually illustrated in Fig. 6a. We analyzed the residuals of the model (Fig. 6b) for normality by running a chi-square test after fitting the residual values to a probability object of a normal distribution. The test did not reject the null hypothesis at the 5% significance level and returned a *p-value* equal to *0.06233*.

5.5.2. Experiment 2: Prediction of the popularity of the "Lipstick on a pig" memetracker phrase

Following the method described in the previous experiment we predicted the evolution of the popularity of the memetracker phrase "Lipstick on a pig". The cumulative number of urls and the acceleration of the adoption rate throughout the

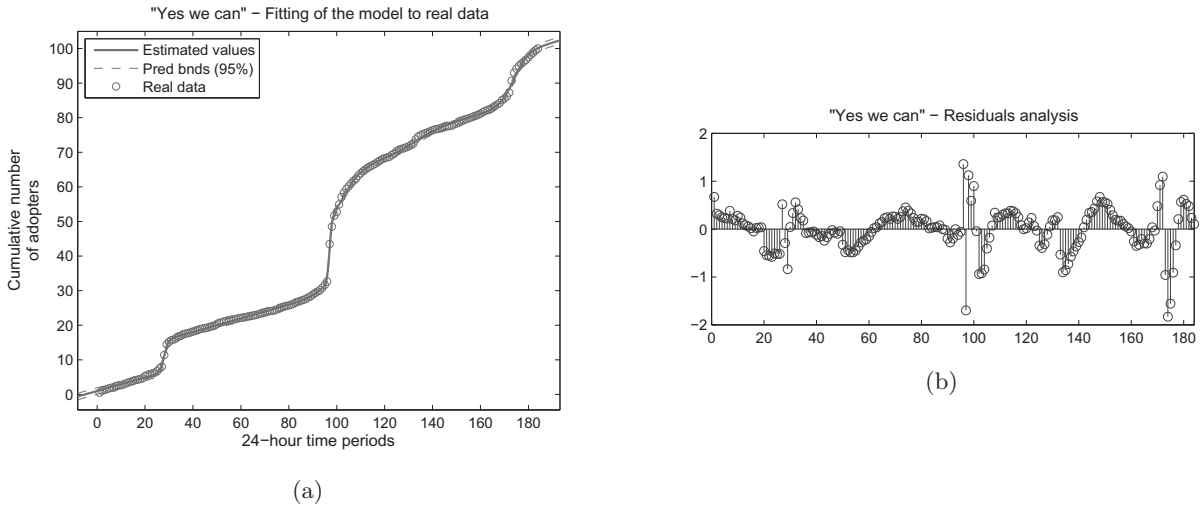

Fig. 6. "Yes we can": (a) Model fitting to real data, (b) Residuals normally distributed up to the 5% significance level, *p-value = 0.06233*.

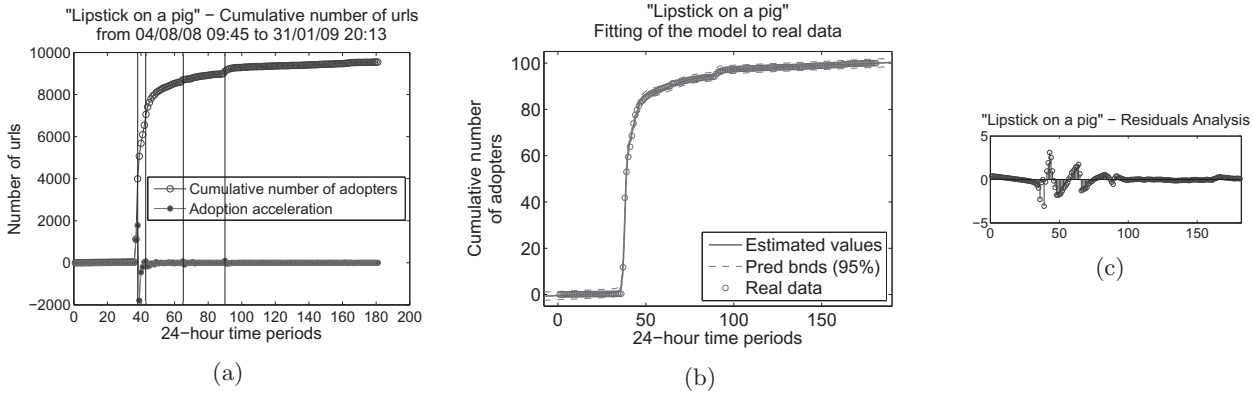

Fig. 7. "Lipstick on a pig": (a) Popularity evolution and adoption acceleration per 24-hour periods, (b) Fitting of the model to the real pattern, (b) Residuals normally distributed up to the 5% significance level, *p-value = 0.1404*.

Table 7
"Lipstick on a pig": Goodness-of-fit of the model to the data (normalized to a scale of 100).

"Lipstick on a pig": Fitting of the model to the data
Model: $N(t) = N(t_0) + P1 * t + (c1/(1 + exp(-a1 * (t - 38)))) + (c2/(1 + exp(-a2 * (t - 43)))) + (c3/(1 + exp(-a3 * (t - 65))))$ $+(c4/(1 + exp(-a4 * (t - 90))))$
Coefficients (95% confidence bounds): a1 = 1.629 (1.474, 1.784), a2 = 0.4752 (0.408, 0.5423), a3 = 0.1799 (0.06433, 0.2955), a4 = 1.293 (-0.9881, 3.575), c1 = 58.37 (56.76, 59.98), c2 = 27.05 (24.58, 29.51), c3 = 6.341 (4.179, 8.503)c4 = 2.719 (1.664, 3.774), $N(t_0)$ = 0.4309 (0.7525, 0.1093), P1 = 0.03277 (0.02577, 0.03977)
Goodness of fit: R-square: 0.9995, Adjusted R-square: 0.9995, RMSE: 0.8861

examined period are depicted in Fig. 7a. The fitting of the final model to the real popularity evolution pattern is illustrated in Fig. 7b and statistically described in Table 7. The residuals of the model – shown in Fig. 7c – are normally distributed. The predictions overlaid on the real values are depicted in Fig. 8. The APE and RSE values of the eight prediction phases are provided in Fig. 9a and b. Table 8 shows the improvement on the performance of the benchmark methods.

5.5.3. Experiment 3: Prediction of the popularity of a flickr photo (id = 14552810)

In this experiment we predicted the popularity of a flickr photo (id = 14552810). The relatively small total number of adopters (∼200) serves as an appropriate test of the effectiveness of the proposed method in producing accurate forecasts even for low popularity online content. The entire popularity growth model and its goodness-of-fit are illustrated in Table 9. The predictions of the model overlaid on the real popularity values are depicted in Fig. 10. The APE and the RSE values of the two prediction phases are shown in Fig. 11a and b. The comparison of the proposed method with the benchmarks

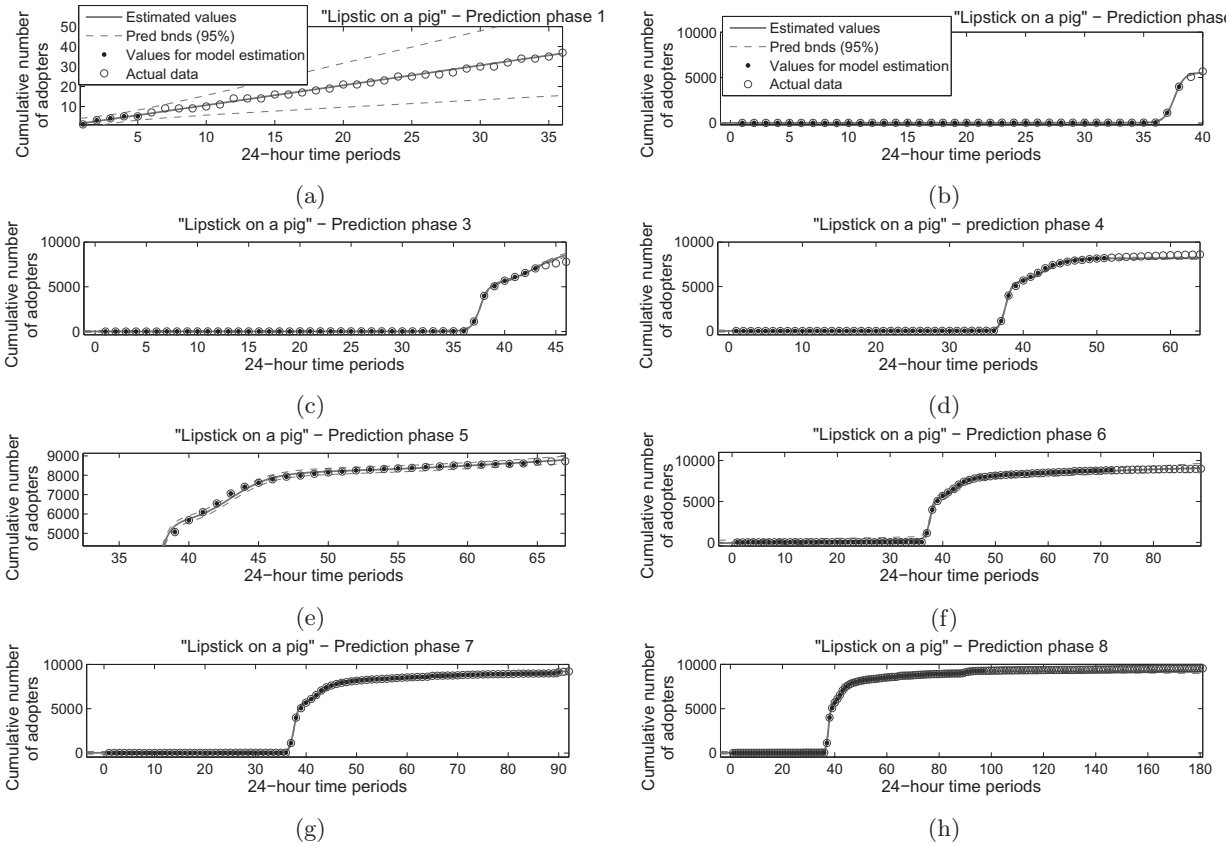

Fig. 8. "Lipstick on a pig": (a)–(h) Prediction phases 1-8. Real values of the cumulative number of adopters (urls) overlaid on the estimated values. The legends of (a) and (b) apply to all the other subfigures.

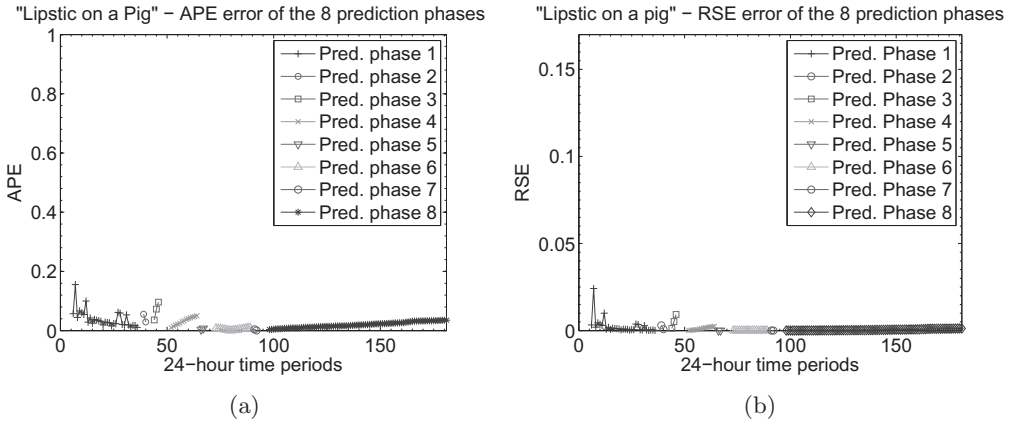

Fig. 9. "Lipstick on a pig": (a) APE values during the eight prediction phases, (b) RSE values during the eight prediction phases.

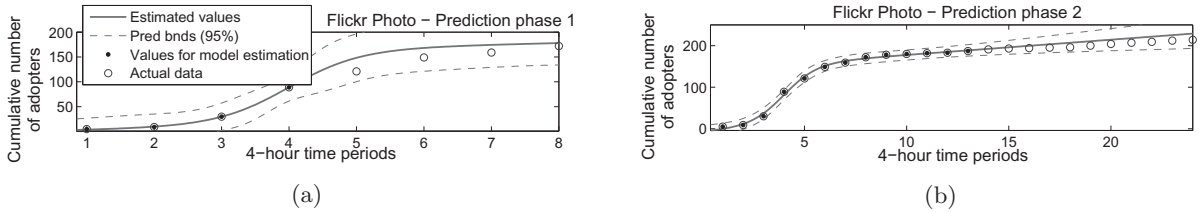

Fig. 10. Flickr photo (id: 14552810)- real popularity values overlaid on the estimated values. (a) Prediction phase 1, (b) Prediction phase 2 (the legend of (a) also applies to (b)).

Table 8
"Lipstick on a pig": Accuracy of the predictions at target time points and comparison with the benchmark methods. The numbers in parentheses indicate the performance of the benchmarks in linear prediction tasks. The performance of the benchmark models was estimated according to the (t_r/t_i) ratio of the prediction task, and the error values reported in Zhao et al. [49], and Pinto et al. [33].

Prediction phase #, type, t_r/t_i	Proposed model (APE & RSE)	SEISMIC model (APE)	S-H model (RSE)	ML model (RSE)
1, linear, 0.1388	0.0108, 0.000117	∼ 0.15	∼ 0.35	∼ 0.33
2, non-linear, 0.5	0.029, 0.000846	N/A (∼ 0.1, linear)	N/A (∼ 0.12, linear)	N/A (∼ 0.1, linear)
3, non-linear, 0.5	0.096, 0.00927	N/A (∼ 0.1, linear)	N/A (∼ 0.12, linear)	N/A (∼ 0.1, linear)
4, linear, 0.28	0.049, 0.00248	∼ 0.12	∼ 0.2	∼ 0.15
5, non-linear, 0.33	0.007, 0.00005116	N/A (∼ 0.12, linear)	N/A (∼ 0.2, linear)	N/A (∼ 0.15, linear)
6, linear, 0.23	0.00855, 0.00007318	∼ 0.125	∼ 0.25	∼ 0.19
7, non-linear, 0.33	0.000443, 0.000000196	N/A (∼ 0.12, linear)	N/A (∼ 0.2, linear)	N/A (∼ 0.15, linear)
8, linear, 0.056)	0.035, 0.00123	∼ 0.2	∼ 0.5	∼ 0.45

Table 9
Flickr photo (id: 14552810): Goodness-of-fit of the model to the data (normalized to a scale of 100).

Flickr photo (id: 14552810): Fitting of the model to the data
Model: $N(t) = N(t_0) + P1 * t + (c1/(1 + exp(-a1 * (t - 4))))$
Coefficients (95% confidence bounds): a1 = 0.9926 (0.9208, 1.064), c1 = 75.4 (73.16, 77.65)
$N(t_0)$ = 7.301 (8.752, 5.849), P1 = 1.071 (1.014, 1.128)
Goodness of fit: R-square: 0.9991, Adjusted R-square: 0.999, RMSE: 0.8428

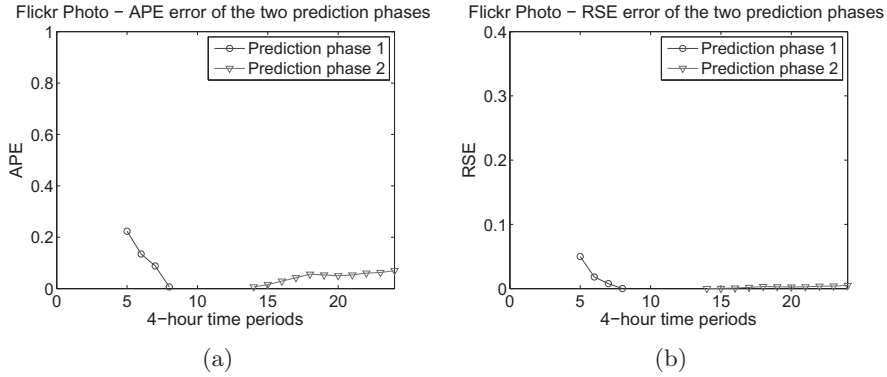

Fig. 11. Flickr photo (id: 14552810): (a) APE values of the two prediction phases, (b) RSE values of the two prediction phases.

Table 10
Flickr photo (id: 14552810): Accuracy of the predictions at target time points and comparison with the benchmark methods. The numbers in parentheses indicate the performance of the benchmarks in linear prediction tasks. The performance of the benchmark models was estimated according to the (t_r/t_i) ratio of the prediction task, and the error values reported in Zhao et al. [49], and Pinto et al. [33].

Prediction phase #, type, t_r/t_i	Proposed model (APE & RSE)	SEISMIC model (APE)	S-H model (RSE)	ML model (RSE)
1, non-linear, 0.5	0.0064, 0.0000409	N/A (∼0.1, linear)	N/A (∼0.12, linear)	N/A (∼0.1, linear)
2, linear, 0.31	0.0698, 0.00488	∼ 0.12	∼ 0.2	∼ 0.15

concerning the accuracy of the predictions at the target time points is provided in Table 10. The fitting of the model to the real data and the normally distributed residuals *(p-value = 0.236)* are depicted in Fig. 12a, and b respectively.

5.5.4. Experiment 4: If we know in advance the inflection point of a sigmoidal adoption phase, how early can we predict the popularity growth pattern and the final number of adopters?

Up to now we have dealt with prediction tasks where no information was available about the inflection point of a non-linear popularity growth phase. Although, such situations are common when the intervention events giving rise to sigmoidal growth patterns cannot be predicted, it is also usual that people are aware of when such events are going to occur. For instance, commercial, social, political or financial important dates might be known in advance, and users start discussing about these issues in social media before these events taking place. As time gets closer to a key date, the adoption rate increases to reach its highest point on the day of the event, and then decays to a stationary pattern. When the key date of an important event is known, the proposed model provides reliable forecasts about the popularity growth of a piece of online content relative to this event well in advance. To demonstrate the predictive capacity of the proposed model in

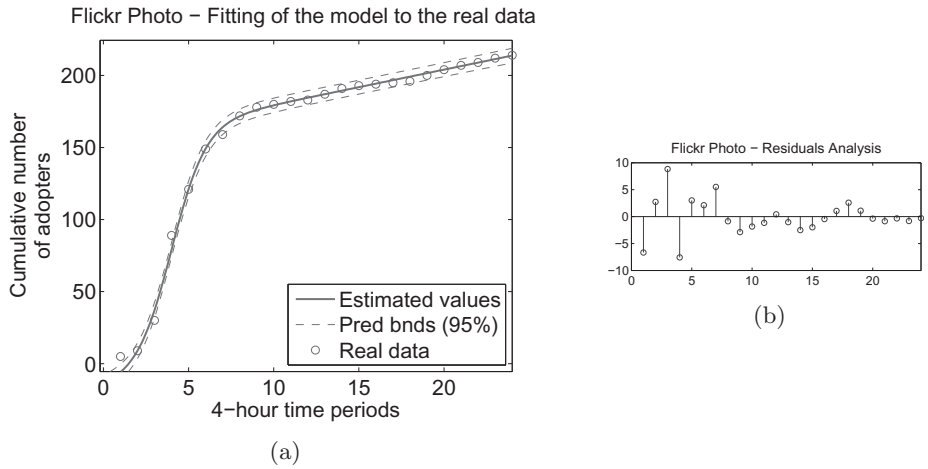

Fig. 12. Flickr photo (id: 14552810): (a) Fitting of the model to real data, (b) Residuals normally distributed up to the 5% significance level, *p-value* = *0.236*.

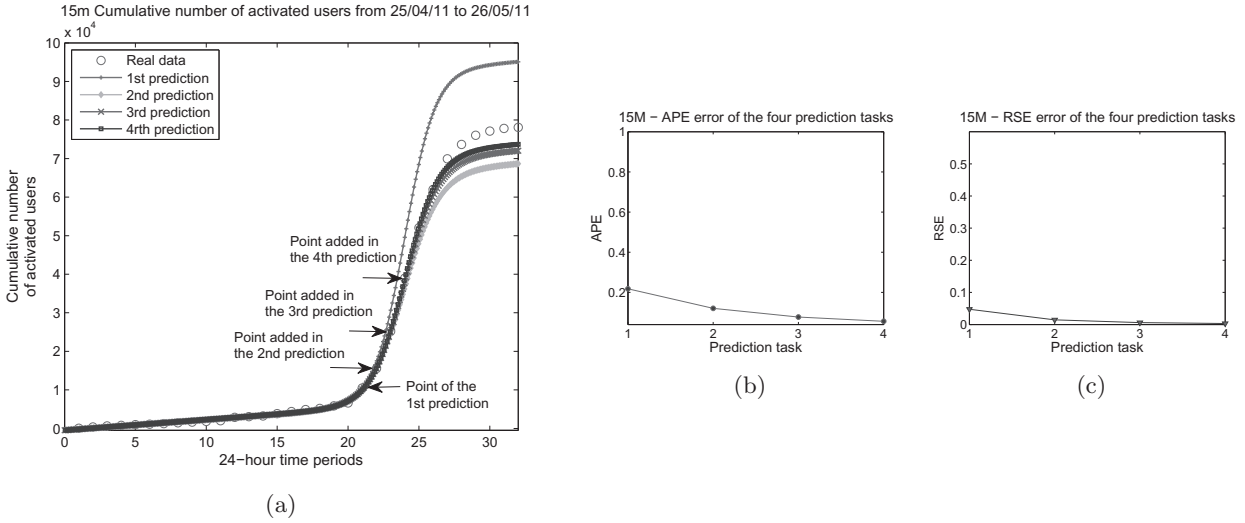

Fig. 13. 15 M online dynamics: (a) Prediction of the total number of activated users assuming that we know the inflection point of the sigmoidal adoption growth. Predictions are generated at four time points indicated by arrows. (b) APE error on the target day of the four prediction tasks. (c) RSE error on the target day of the four prediction tasks.

such scenarios, we performed an experiment using the daily growth of the number of Twitter users who posted at least one message about the 15 M movement in Spain in the period from April 25, 2011 to May 26, 2011. In this experiment we assumed that we knew that the inflection point of the sigmoidal popularity pattern would be on day 24 from the 32 days of the analyzed period. The prediction question we aim to answer can be stated as follows: *Suppose that the cumulative number of adopters per day is available, how early can we predict the total number of adopters on a target day, if we know that the infection point of the sigmoidal popularity growth is going to be on day 24?*

We examined four prediction scenarios with forecasts generated on day 21, 22, 23, and 24 respectively. The prediction results are illustrated in Fig. 13a. The APE and RSE values of the predictions about the total number of adopters indicate an excellent performance even from the 1st prediction, thus outperforming the benchmark methods which report larger errors in simpler, linear prediction tasks. We also mention that the benchmark methods do not cover the prediction scenario of this example. In this experiment we used the predictive model described by the equation $N(t) = N(t_0) + P_1 \cdot t + C/(1 + e^{-a(t-24)})$.

The APE and RSE errors of the four prediction tasks regarding the total number of adopters on the target day t_{32} are depicted in Fig. 13b and c. Considering that similar adoption patterns appear in various socioeconomic contexts, such as financial markets, advertising and cultural events, the proposed method has the potential to predict the popularity growth in such settings with the same success.

Table 11
7/30 prediction test: Goodness-of-fit of the model to the data.

7/30 prediction test: Fitting of the model to the data
Model: $N(t) = P1 * t + (c/(1 + exp(-a * (t - 3))))$
Coefficients (95% confidence bounds): a = 2.993 (0.1467, 6.133), c = 1289 (765.6, 1812), P1 = 42.79 (-42.45, 128)
Goodness of fit: R-square: 0.9971, Adjusted R-square: 0.9957, RMSE normalized to a scale of 100: 1.583

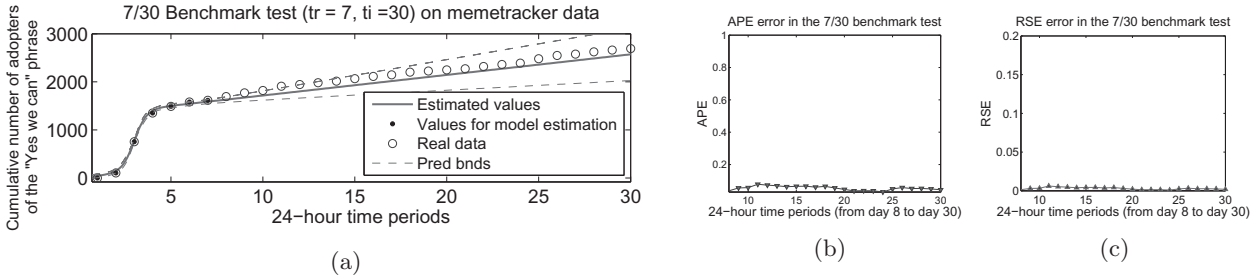

Fig. 14. 7/30 benchmark test: (a) Fitting of the predictions to the real data, (b) APE values, (c) RSE values.

Table 12
Comparison of the target day predictions in the 7/30 benchmark test. The error values of the S-H, ML, and MRBF models were obtained from Pinto et al. [33]. The APE value of the SEISMIC model was obtained from Zhao et al. [49].

Prediction task	Proposed model (APE & RSE)	S-H model (median RSE)	ML model (median RSE)	MRBF model (median RSE)	SEISMIC model (median APE)
$t_r/t_i = 7/30$	*RSE*: 0.00193, *APE*: 0.0439	Random dataset: 0.2382 ± 0.0038, Top dataset: 0.2121 ± 0.0074	Random dataset: 0.2022 ± 0.0043, Top dataset: 0.1837 ± 0.0081	Random dataset: 0.1892 ± 0.0032, Top dataset: 0.1723 ± 0.0071	~ 0.12

5.5.5. Experiment 5: 7/30 test against the S-H, ML, and MRBF benchmark models

The S-H, ML and MRBF models predict the popularity of online content using the correlation coefficients between the initial and future popularity at various time points. The problem addressed by these models is essentially the popularity prediction of a piece of online content after an initial popularity burst. The prediction method proposed in this study has demonstrated its capacity to generate popularity forecasts not only in situations where the initial explosive phase is followed by a linear popularity growth, but also in more complex cases where a linear popularity evolution phase is interrupted by non-linear phases caused by intervention events. However, in order to compare the performance of the proposed method with that of the S-H, ML and MRBF methods in a common prediction scenario, we extracted from the popularity growth pattern of the memetracker phrase "Yes we can", the part from the time point t_{27} to the time point t_{57}, which resembles the case addressed by the aforementioned benchmark models. From the popularity values we subtracted the value of the first data point, so the popularity growth time-series starts from zero, as if it were the start point of the popularity evolution of the "Yes we can" phrase. Using this time-series we run the 7/30 benchmark test of the study by Pinto et al. [33]. This test uses information about the total number of adopters of a piece of online content in the first seven days, in order to predict its popularity on the 30th day ($t_r/t_i = 7/30$). Using a model consisting of a logistic and a 1st-degree polynomial term (see Table 11), we accurately predicted the popularity evolution in a way that substantially outperformed the benchmark models. Fig. 14a shows the real values overlaid on the predictions of the model. The RSE and APE value of the popularity forecast on the target day (t_{30}) generated by the proposed model, along with the corresponding values of the benchmark models, are provided in Table 12. It can be seen that the improvement on the performance of the benchmarks is remarkable. Figure 14b and 14c show the APE and RSE values during the entire prediction horizon.

5.5.6. Experiment 6: Application of the method to predictions in a timescale of minutes

The proposed method is equally effective when using higher time resolutions in the adoption rate. To demonstrate how the method performs using a timescale of minutes, we predicted the size of the first extensive, non-linear adoption cascade in the propagation of the #suarez hashtag. Fig. 15a shows how the number of new adopters and the number of tweets containing the #suarez hashtag evolved over time. Fig. 15b shows the acceleration of the adoption rate per minute before and after the first sigmoidal growth phase. The similarity of the fluctuations of the adoption acceleration with seismograms is remarkable. Probably this similarity could be further investigated in relation to the analogy between the dynamics of information cascades and that of earthquakes. Fig. 15c provides a closer view on the examined time period and depicts the start t_s, the inflection t_{infl}, and the end t_e time points of a non-linear adoption cascade, whose size we are going to predict.

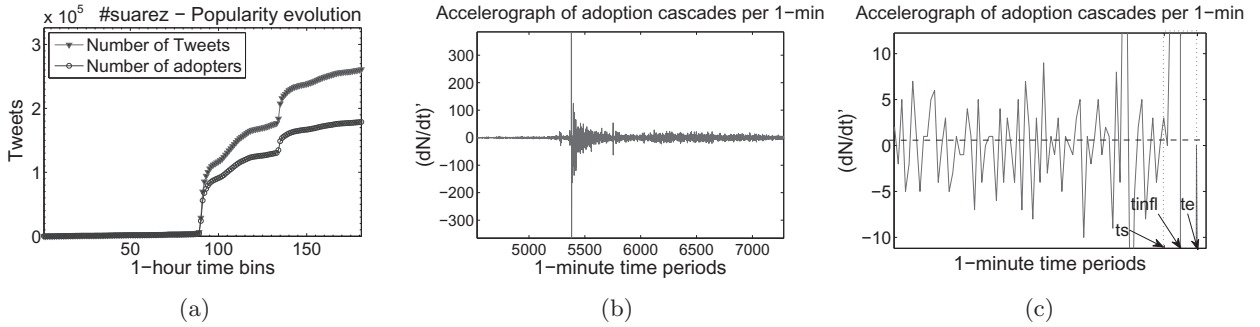

Fig. 15. (a) Growth of the number of hashtag adopters and Twitter messages containing the #suarez hashtag. (b) Adoption acceleration per minute before and after the 1st non-linear growth phase. (c) Closer view on the examined period. Note: The similarity of the fluctuations of the acceleration rate with seismograph records is remarkable.

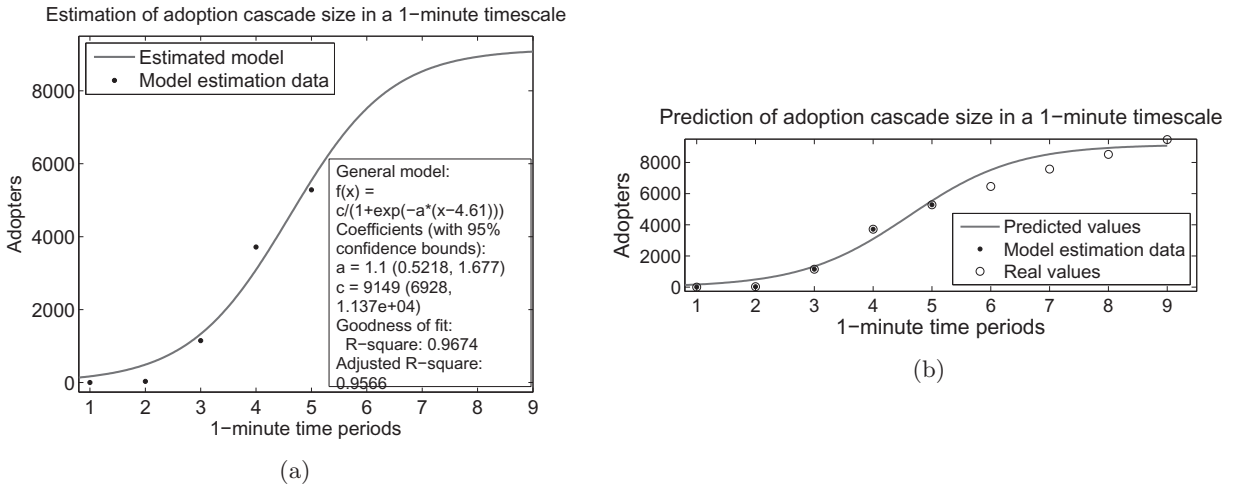

Fig. 16. (a) Model estimation. (b) Predicted values overlaid on the real ones.

Using the time point t_{infl} and the number of adopters $N(t)$ from t_s to the first time point $t_i >= t_{infl}$, we estimated the parameters of the logistic adoption growth and accurately predicted the size of the non-linear cascade. Figure 16a depicts the predictive model and the popularity forecasts. Figure 16b visually illustrates the fitting of the predicted popularity values to the real ones. Both in Fig. 16a and in 16b the origins of *y-axis* are set to *0*, so the values along *y-axis* correspond to the size of the cascade. The *RMSE* of the popularity predictions after normalizing the predicted and the real popularity values to a reference scale of 100 was estimated at *3.72*, thus statistically confirming the high accuracy of the predictions.

Comparison with a benchmark model. In order to provide quantitative evidence of the improvement on the accuracy of the predictions concerning the sizes of adoption bursts, we compared the proposed method with the method by Kong et al. [21]. This method estimates the logarithm of the sizes of Twitter activity bursts in relation to a specific hashtag. The prediction problem that we investigate is similar, since the growth pattern of Twitter activity is equivalent to the growth pattern of the adopters of a hashtag, as shown in Fig. 15a. We mention that the proposed method predicts the real size of adoption bursts, which is a more difficult task than predicting the logarithm of their sizes. However, for benchmarking purposes we compared the RMSE of the logarithms of the predictions generated by the proposed method with the RMSE value reported by Kong et al. [21] in prediction tasks of a 1-minute time resolution. After calculating the RMSE of the proposed method using the normalized values of the logarithms of the predictions in the scale 0-15 of the reference study, we measured an *86.23%* improvement (0.1445 vs. 1.05).

5.5.7. Experiment 7: Prediction of the number of adopters of a low-popularity, fleeting Twitter hashtag

In this experiment we predicted the number of adopters of the #100thingsiwant hashtag obtained from the dataset of the study by Weng et al. [43]. The data covers a period of ~ 97 hours from 15/04/12 12:55 to 19/04/12 13:36. This hashtag serves as an ideal test case for the evaluation of the capacity of the proposed method to predict the adoption growth of fleeting, low popularity online content. As shown in Fig. 17a the popularity evolution pattern of this hashtag comprises one non-linear growth period followed by a short-lived linear period. We predicted the popularity of the #100thingsiwant

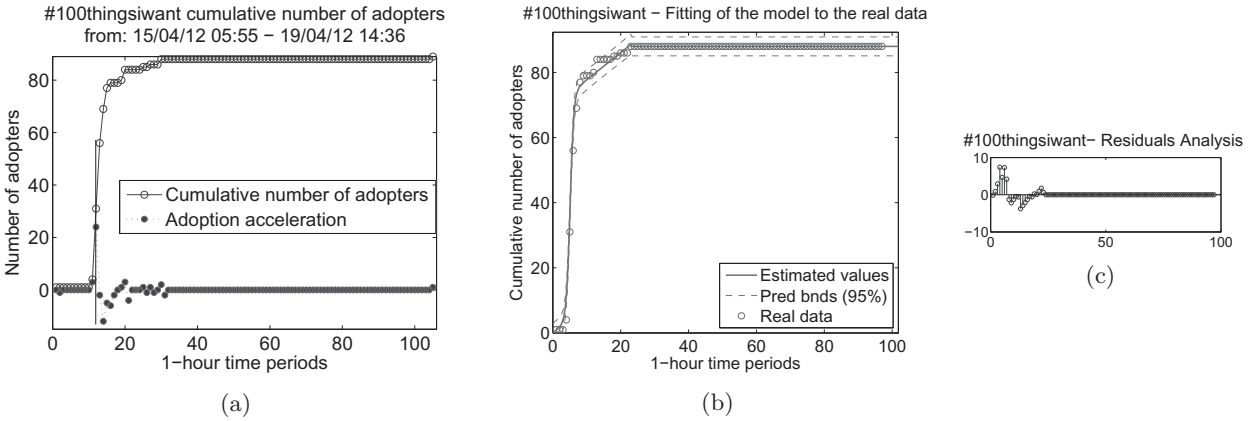

Fig. 17. #100thingsiwant: (a) Popularity evolution and acceleration of the adoption rate per 1-hour periods. (b) Fitting of the model to the real data. (c) The residuals are normally distributed up to the 5% significance level, *p-value = 0.0571*.

Table 13
#100thingsiwant: Goodness-of-fit of the model to the data.

#100thingsiwant: Fitting of the model to the data
Model: $for\, t \le 24 :\ N(t) = (c1/(1 + exp(-a1 * (t - 5)))) + P1 * t_{t \le 24}$ $for\, t > 24 :\ N(t) = (c1/(1 + exp(-a1 * (t - 5)))) + P1 * t_{t \le 24} + P2 * (t_{t>24} - t_{24})$ Coefficients (95% confidence bounds): a1 = 1.83 (1.576, 2.084), c1 = 69.7 (67.45, 71.95) P1 = 0.8094 (0.6724, 0.9464), P2 = 1.685e-18 (-0.005289, 0.005289) Goodness of fit: R-square: 0.9937, Adjusted R-square: 0.9935, RMSE: 1.469 (using a reference scale of 100 for the model estimations and the real data the RMSE was estimated at 1.669)

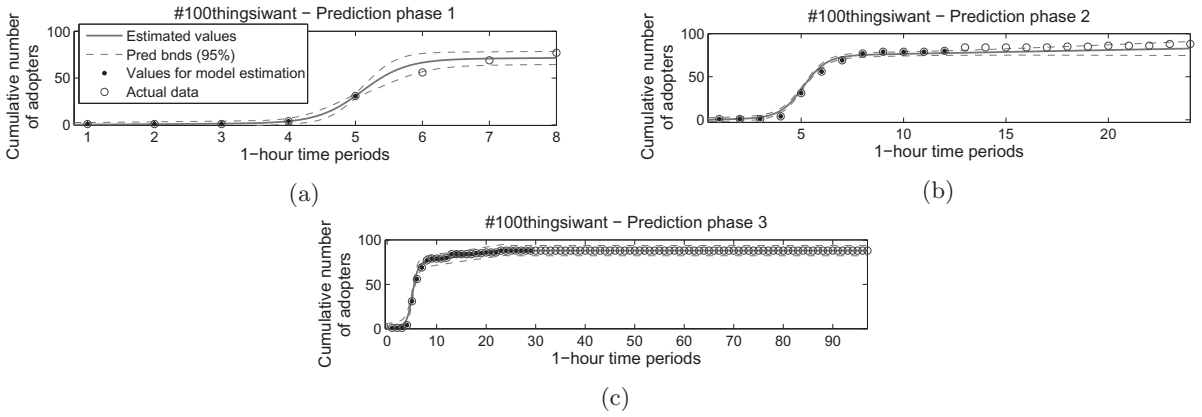

Fig. 18. #100thingsiwant: Real popularity values overlaid on the estimated values. (a) Prediction phase 1, (b) Prediction phase 2, (c) Prediction phase 3. The legend of (a) also applies to (b) and (c).

through three prediction phases. In the first phase by estimating the parameters of the sigmoidal adoption pattern, we predicted the evolution of the non-linear popularity growth. In the second phase we estimated the stationary adoption rate *P1* thereby predicting the popularity during the short linear growth phase. Finally, in the third prediction phase after detecting that the adoption acceleration was zero from t_{25} to t_{29}, we re-estimated the linear growth trend by using the total number of adopters from t_{25} to t_{29}. The adoption rate *P2* was estimated at *1.685e-18*, that is, zero. A zero growth rate implies the termination of the popularity growth of the #100thingsiwant hashtag, and therefore the model correctly predicted that the total number of adopters was not going to change.

The goodness-of-fit of the entire model to the real data is illustrated in Table 13, while a visual representation of the model fit to the real popularity pattern is provided in Fig. 17b. The residuals of the model (Fig. 17c) are normally distributed up the 5% significance level (*p-value = 0.0571*). The results of the three prediction phases are depicted in Fig. 18. The APE and RSE values of the predictions are illustrated in Figs. 19a and b, respectively. The evaluation of the performance of the proposed method against the benchmarks is provided in Table 14.

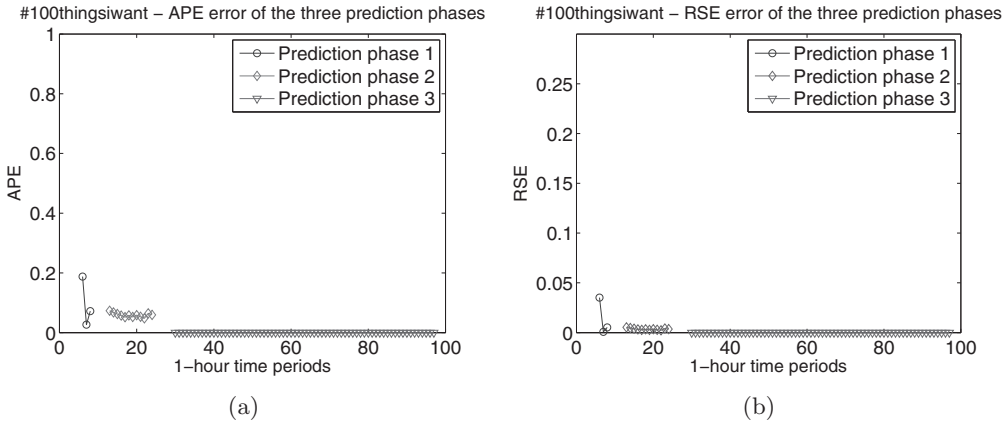

Fig. 19. #100thingsiwant: (a) APE values of the three prediction phases, (b) RSE values of the three prediction phases.

Table 14
#100thingsiwant: Comparison of the accuracy of the predictions at target time points. The numbers in parentheses indicate the performance of the benchmarks in linear prediction tasks. The performance of the benchmark models was estimated according to the (t_r/t_i) ratio of the prediction task, and the error values reported in Zhao et al. [49] and Pinto et al. [33].

Prediction phase #, type, t_r/t_i	Proposed model (APE & RSE)	SEISMIC model (APE)	S-H model (RSE)	ML model (RSE)
1, non-linear, 0.625	0.072, 0.0052	N/A ($\sim$ 0.09, linear)	N/A ($\sim$ 0.07, linear)	N/A ($\sim$ 0.05, linear)
2, linear, 0.25	0.0594, 0.0035	$\sim$ 0.125	$\sim$ 0.23	$\sim$ 0.18
3, linear, 0.068	0, 0	N/A prediction error in non-persistent online content	N/A prediction error in non-persistent online content	N/A prediction error in non-persistent online content

6. Algorithmic implementation of the popularity prediction method

6.1. Algorithm of the prediction method

In what follows we formulate the prediction method into an algorithm aiming to the real time prediction of the popularity of online content. A list of the symbols used in the algorithm is provided in Table 15.

Prediction method algorithm

Input: Number of new adopters per time period, $n(t_i),\ for\, i = 0, 1, 2, ..., T$
Initialization: $N(t_0) \leftarrow N_0;\ \gamma(t_0) \leftarrow 0;\ n(t_0) \leftarrow n_0;\ k \leftarrow 0;\ f \leftarrow 0;\ NL_{in} \leftarrow false;\ Z_{accel} \leftarrow 0;$
$\Delta \leftarrow user\ defined\ value$
$L \leftarrow user\ defined\ value$
$dp_s \leftarrow user\ defined\ value$
for $i = 1$ to T **do**
  $N(t_i) \leftarrow N(t_{i-1}) + n(t_i);$
  $\gamma(t_i) \leftarrow n(t_i) - n(t_{i-1});$
  $/*$ *Construction of the linear parts of the model* $*/$
  $/*$ *Predictions during stationary periods* $*/$
  if $-\Delta \leq \gamma(t_i) \leq \Delta$ **then**
    if $(k = 0)$ **or** $(i \bmod L = 0)$ **or** $(NL_{ex} = true)$ **then**
      $S_{start} \leftarrow i;$
      $Smdl_{est} \leftarrow false;$
    end if
    if $\gamma(t_i) = 0$ **then**
      $Z_{accel} \leftarrow Z_{accel} + 1;$
      $/*$ *Detection of the popularity growth termination* $*/$
      if $Z_{accel} = dp_s$ **and** $([\gamma(t_i), \gamma(t_{i-1}), ..., \gamma(t_{i-dp_s+1})] = 0)$ **then**
        $S_{start} \leftarrow i - dp_s + 1;$
        $Z_{accel} \leftarrow 0;$
        $Smdl_{est} \leftarrow false;$

end if
end if
if $i = S_{start} + dp_s - 1$ **then**
$newS \leftarrow true$;
$k \leftarrow k + 1$;
$S(k, 1) \leftarrow S_{start}$;
if $k \geq 2$ **then**
$S(k-1, 2) \leftarrow S_{start} - 1$;
end if
if NL_{ex} **then**
$NL_{ex} \leftarrow false$;
end if
$\mathbf{P_k^{S(k,1):i}} = \arg\min_{\mathbf{P_k'}} \text{llsq}(\text{N}(\text{t}_{\text{S(k,1)}}), \ldots \text{N}(\text{t}_\text{i}), \mathbf{P_k'})$; /* *Parameter estimation* */
$newS \leftarrow false$;
$Smdl_{est} \leftarrow true$;
if $(f = 0)$ **and** $(k > 0)$ **then**
$mdl \leftarrow \sum_{j=1}^{k-1} P_j^{S(j,1):S(j,2)} * (S(j, 2) - S(j, 1)) + P_k^{S(k,1):i} * (i - S(k, 1))$;
else
if $(k > 0)$ **and** $(f > 0)$ **then**
$mdl \leftarrow \sum_{j=1}^{k-1} P_j^{S(j,1):S(j,2)} * (S(j, 2) - S(j, 1)) + \sum_{l=1}^{f} \frac{C_l}{1+e^{(-a_l*(i-t_{infl}^l))}} + P_k^{S(k,1):i} * (i - S(k, 1))$;
end if
end if
end if
if $Smdl_{est}$ **then**
$t_p \leftarrow$ *user defined value*;
$\widehat{N(t_p)} \leftarrow N(t_i) + P_k^{S(k,1):i} * (t_p - i)$; /* *Generation of forecasts* */
return $\widehat{N(t_p)}$;
end if
end if

/* *Construction of the non − linear parts of the model* */
/* *Predictions during the non − linear periods of the adoption growth* */

if $(\gamma(t_i) > \Delta)$ **and** $(\textbf{not}\ NL_{in})$ **then**
$f \leftarrow f + 1$;
$NL(f, 1) \leftarrow i - 2$;
$NL_{in} \leftarrow true$;
$NLmdl_{est} \leftarrow false$;
$D_f^{expl} \leftarrow i - NL(f, 1) + 1$;
end if
if $(\gamma(t_i) > 0)$ **and** NL_{in} **then**
$D_f^{expl} \leftarrow D_f^{expl} + 1$; /* *Measuring the duration of an exponential growth phase* */
end if
if $(\gamma(t_i) < 0)$ **and** NL_{in} **then**
$t_{infl}^f \leftarrow i - 1$; /* *or the point where the line* $(\gamma(t_{i-1}), \gamma(t_i))$ *crosses the zero axis* */
$NL(f, 2) \leftarrow t_{infl}^f + D_f^{expl}$;
$\{\mathbf{C_f}, \mathbf{a_f}\} = \arg\min_{\mathbf{C_f'}, \mathbf{a_f'}} \text{nllsq}(\text{N(f, 1)}, \ldots, \text{N}(\text{t}_{\text{infl}}^\text{f}), \text{t}_{\text{infl}}^\text{f}, \mathbf{C_f'}, \mathbf{a_f'})$; /* *Parameter estimation* */
$NL_{in} \leftarrow false$;
$NLmdl_{est} \leftarrow true$;
if $k > 0$ **then**
$mdl \leftarrow \sum_{j=1}^{k-1} P_j^{S(j,1):S(j,2)} * (S(j, 2) - S(j, 1)) + P_k^{S(k,1):i} * (i - S(k, 1)) + \sum_{l=1}^{f} \frac{C_l}{1+e^{(-a_l*(i-t_{infl}^l))}}$;
else
$mdl \leftarrow p_{intcpt} + \sum_{l=1}^{f} \frac{C_l}{1+e^{(-a_l*(i-t_{infl}^l))}}$;
end if
$\widehat{N(t_{NL(f,2)})} \leftarrow N(t_{NL(f,1)}) + \frac{C_f}{1+e^{(-a_f*(t_{NL(f,2)}-t_{infl}^f))}}$; /* *Generation of forecasts* */

76: **return** $\widehat{N(t_{NL(f,2)})}$;
77: **end if**
78: **if** $(i \leq NL(f,2))$ **and** $NLmdl_{est}$ **then**
79: $t_{nl} \leftarrow user\,defined\,value \leq NL(f,2)$;
80: $\widehat{N(t_{nl})} \leftarrow N(t_{NL(f,1)}) + \frac{C_f}{1+e^{(-a_f*(t_{nl}-t^f_{infl}))}}$;
81: **return** $\widehat{N(t_{nl})}$;
82: **else**
83: **if** $NLmdl_{est}$ **then**
84: $NL_{ex} \leftarrow true$;
85: **end if**
86: **end if**
87: **end for**

6.2. Description of the algorithm of the prediction method

Initialization part. The input information to the popularity prediction algorithm is the number of new adopters $n(t_i)$ per time period t_i, where i represents units of temporal resolution (e.g. time bins of minutes, hours, days). This is described in line *1* of the algorithm. The process runs for T time periods. Before generating forecasts we initialize the total number of adopters $N(t)$, the acceleration rate $\gamma(t_i)$, the number of new adopters $n(t_0)$ at t_0, the number of stationary and non-stationary periods – k and f respectively – the flag NL_{in}, and the counter Z_{accel} of the time periods with zero adoption acceleration. The initialization process is described in line *2*. The flag NL_{in} shows whether the adoption process undergoes a non-linear phase or not. A complete description of the symbols of the algorithm is provided in Table 15. To generate forecasts we also need to define the adoption acceleration threshold Δ signifying the passage of an adoption process from stationarity to non-stationarity and vice versa. The optimal value of Δ results from a trade off between prediction accuracy and frequency of the re-estimation of the model parameters. The parameter L in line *4* represents the number of time periods after which the current linear growth term of the predictive model has to be re-estimated in order to adjust to possible variations in the stationary adoption trend. The higher the value of Δ is, the lower the L should be, so the process can compensate for the error induced by relatively large values of Δ. The parameter dp_s represents the number of time points that are used in the estimation of the linear trend. In the examples presented in this study, dp_s was set to 4 or 5.

Processing of input information. After the initialization stage, the algorithm proceeds as follows. For each time period from *0* to T we calculate the total number of adopters $N(t_i)$ (line *7*). We also calculate the acceleration of the adoption rate $\gamma(t_i)$ (line *8*). This information is used at later stages of the process dealing with the construction of the predictive model, the estimation of the values of its parameters from the data, and the generation of popularity forecasts.

Detection of linear growth phases. Whether the online content adoption acceleration $\gamma(t_i)$ is within the range $[-\Delta, \Delta]$ is checked at each time period t_i (line *11*). If this is true the process examines three additional conditions. First, whether no linear phases have been already detected *(k=0)*; second, whether the number of time periods elapsed from the beginning of the process is a multiple of the re-estimation time interval L *(I mod L = 0)*; and third, whether the process has exited from a non-linear adoption period. In all these cases the method should estimate a linear popularity growth trend in order to start producing forecasts since the online content adoption process undergoes a stationary phase. If any of these conditions is true, the process stores the current time point t_i to the variable S_{start} (line *13*), so it can recall it at a later step. Also, the flag $Smdl_{est}$ indicating whether a linear model has been already estimated or not, is set to false (line *14*). However, there is another special case in which the method should estimate a linear growth trend. This case relates to the detection of the termination of the popularity growth of a piece of online content. The algorithm handles this scenario from line *16* to line *24*. In particular, the process examines whether the adoption acceleration of the current time period is equal to zero. If this condition is true the counter Z_{accel} increases by one starting from zero. When this counter becomes equal to dp_s and the adoption acceleration of dp_s consecutive time periods is equal to zero, then it is most likely that the popularity growth of the monitored online content has finished. In this case a new linear trend of zero growth should be estimated, so that the popularity forecasts can account for the fact that no other adoptions are likely to occur. The start point of the new stationary phase of zero popularity growth is set to $i - dp_s + 1$. The process also resets the counter Z_{accel} and sets the flag $Smdl_{est}$ to false, since the new linear growth trend has not yet been estimated. For the estimation of a linear growth trend we need a number of data points. Whether an adequate number of such points is available or not is examined in the *"if"* statement of line *25*. Since we need dp_s data points for the estimation of a linear growth trend including the start point (S_{start}) of the ongoing linear phase, the current time period should be equal to $S_{start} + dp_s - 1$. When this happens the flag *newS* is set to true (line *26*), the number of linear phases k increases by one (line *27*), and the start point *S(k,1)* of the k^{th} stationary phase is set to S_{start} (line *28*). Notice that the condition of line *25* will be true when the process successfully exits from the "if" statement of line *19* dealing with the detection of the termination of an adoption process. If we have already detected more than two stationary phases, then the end point of the previous one *S(k-1,2)* is set to the value $S_{start} - 1$, as shown in

Table 15
Description of symbols of the algorithm of the prediction method.

Symbol	Description
$N(t_i)$	Cumulative number of adopters at the time point t_i
$n(t_i)$	Number of new adopters at the time point t_i
$\gamma(t_i)$	Acceleration of the adoption rate at the time point t_i
T	Length (number of time periods) of the prediction process
k	Number of linear adoption phases
f	Number of non-linear adoption phases
NL_{in}	Flag indicating that the adoption process is in a non-linear phase
NL_{ex}	Flag indicating that the adoption process exited a non-linear phase
dp_s	Number of data points for the estimation of a linear term of the model
Z_{accel}	Counter of the time points at which the acceleration of the adoption rate is equal to zero
Δ	Acceleration rate threshold indicating the transition from a stationary to a non-stationary adoption phase
L	The number of time periods after which the current linear growth rate has to be re-estimated
S_{start}	The time point at which a new stationary adoption phase starts
$Smdl_{est}$	Flag indicating that a new linear term of the model has been estimated
newS	Flag indicating that a new stationary adoption phase has started
S(k,1:2)	A *(k,2)* array where the start *S(k,1)* and the end time point *S(k,2)* of the k stationary phases are stored
$P_k^{S(k,1):i}$	Adoption rate of the ongoing stationary phase k, started at $t_{S(k,\,1)}$
$P_j^{S(j,1):S(j,2)}$	Adoption rate of the j^{th} stationary phase, started at $t_{S(j,\,1)}$, and ended at $t_{S(j,\,2)}$
t_p	A future time point for which a forecast about the total number of adopters is required
t_{nl}	A future time point before the expected end of a non-linear phase for which a forecast about the total number of adopters is required
$\widehat{N(t_p)}$	Estimated total number of adopters at the time point t_p
D_f^{expl}	Duration of the exponential increase phase of a non-linear popularity growth period
NL(f,1:2)	An *(f,2)* array where the start *NL(f,1)* and the end time point of the f^{th} non-linear adoption period are stored
t_{infl}^f	Inflection point of the f^{th} sigmoidal popularity growth phase
C_f	Size of the f^{th} non-linear cascade
a_f	Slope of the f^{th} sigmoidal popularity growth phase
mdl	Model describing the actual popularity growth up the current time point t_i
$NLmdl_{est}$	Flag indicating that a non-linear term of the model has been estimated
$N(t_i)$	Number of new adopters at the time point t_i
p_{intcpt}	Intercept point of the model when the adoption process starts with a non-linear phase

line *30*. The condition of line *32* examines whether the adoption process entered the current stationary phase after exiting from a non-linear adoption period. In such a case the flag NL_{ex} is set to false, so the process can use it again when a future non-linear adoption phase will have finished.

Estimation of a 1st-degree polynomial part of the model. At this stage of the process dp_s data points are available for the estimation of a 1st-degree polynomial part of the predictive model. Such a part features the form $N(t) = P_k \cdot t$, where P_k denotes a stationary adoption trend. These points are used as input to a linear least squares optimization process (line *35*) producing an estimate P_k' of the stationary adoption rate minimizing the quantity $S = \sum_{i=1}^{n} r_i^2$, that is, the sum of the squares of the residuals, given by $r_j = N(t_j) - \widehat{N(t_j)}$, where $j = S(k,1), \ldots, i$ is the observation number, $N(t_j)$ is the actual cumulative number of adopters, and $\widehat{N(t_j)}$ is the corresponding value estimated by the fitted model. After the construction of a linear part of the predictive model the flag *newS* indicating that the process is in the phase of estimating a new stationary adoption trend is set to false (line *36*). Also the flag $Smdl_{est}$ is set to true (line *37*), thus showing that a new linear part of the model has been estimated, and therefore it can be used in the prediction of the online content popularity growth.

Construction of the mathematical expression of the model. To construct a mathematical formula for the model, the algorithm examines the number f of the non-linear, and the number k of the linear parts of the model up to the current time period t_i. If the number of the non-linear parts is zero, and the number of stationary parts is greater than zero – the condition in line *38* - then the model features the form described in line *39*. It can be seen that it consists of all the previous stationary phases, included in the summation term, and the current stationary phase. While all the previous stationary phases have a start *S(j,1)* and an end point *S(j,2)*, the current stationary phase has not yet an end point as it is still ongoing. When the process has already encountered both stationary and non-stationary phases, in other words when $k > 0$ and $f > 0$, the condition of the *"if"* statement in line *41* becomes true and the model takes the form described in line *42*.

Generation of forecasts during linear growth phases. After estimating the parameters of the predictive model, the flag $Smdl_{est}$ becomes true and the process is ready to generate forecasts. The readiness state is examined in the condition of the *"if"* statement in line *46*. If the process is ready the user defines a future time period t_p as shown in line *47*. Then in line *48* the model produces an estimate $\widehat{N(t_p)}$ of the popularity at the time period t_p by adding to the current popularity $N(t_i)$, the extrapolation of the estimated stationary trend P_k for a number of time periods equal to the difference between t_p and

the current time period t_i. The estimated value is returned to the user in line *49*. As long as the predictive model is not re-estimated the prediction is valid.

Detection of non-linear growth phases. In line *54* the algorithm examines whether the acceleration of the adoption rate $\gamma(t_i)$ has exceeded the stationarity threshold Δ, and whether the process has not already been within a non-linear adoption period ($NL_{in} = false$). The second condition is necessary because the acceleration of the adoption rate might be higher than Δ, but its prior value might also be higher than Δ, thus meaning that the process had already entered a non-linear adoption period. In such a case the process should not consider an acceleration rate higher than Δ as a signal of entering a new non-linear adoption period. When the process successfully passes the conditions of the *"if"* statement in line *54*, the number f of the non-linear phases increases by one as shown in line *55*. Also, the start of the non-linear period *NL(f,1)* is set to two time points before the current one (line *56*), so that the lower asymptote of the sigmoidal adoption pattern is considered in the estimation of the parameters of the non-linear part of the model. In line *57* the flag NL_{in} indicating that the process is already within a non-linear adoption phase is set to *"true"*. In line *58* the flag $NLmdl_{est}$ indicating that a new non-linear part of the model has been estimated is set to false, because the process needs some additional information before estimating the parameters of the new non-linear term of the model. In line *59* the counter D_f^{expl} measuring the duration of the exponential increase phase in a non-linear popularity growth period is set to the value $i - NL(f, 1) + 1$, so it captures all the time points in the interval *[NL(f,1), i]*, that is, from the start of a non-linear adoption period up to the current time point t_i. In line *61* the process examines whether the acceleration rate is positive and whether the adoption process is in a non-linear phase ($NL_{in} = true$). The second condition is necessary because the acceleration rate can also be positive (or negative) during stationary periods, as it presents small fluctuations around zero. If both conditions are true the duration of the exponential phase increases by one time period as shown in line *62*. The next *"if"* statement in line *64* examines whether the acceleration rate has changed trend from positive to negative, and whether the adoption process has already been in a non-stationary phase ($NL_{in} = true$). The second condition is necessary for the reason explained before, that is, to ensure that the changes in the acceleration trend do not occur within a stationary phase. If both conditions are true, the process sets the inflection point t_{infl}^f of the f^{th} non-linear adoption period to the value *i - 1* (line *65*), and the end point *NL(f,2)* of the f^{th} non-linear phase is set to the value $t_{infl}^f + D_f^{expl}$ (line *66*), since the inflection point of a logistic growth pattern is in the middle of the sigmoidal curve. We mention that another estimate of the inflection point t_{infl}^f is the time point where the line $(\gamma(t_{i-1}), \gamma(t_i))$ crosses the zero axis. In most of our experiments the former was the case.

Estimation of a non-linear part of the model. Having estimated the inflection point of the f^{th} non-linear part of the model, and having available the data points of the cumulative number of adopters from its start point $t_{NL(f,\,1)}$ to its inflection point t_{infl}^f, we can estimate the size of the nonlinear cascade, and the steepness of the sigmoidal adoption curve by entering these values into a non-linear least squares optimization method as shown in line *67*. The optimization process should start from a high initial C_f value, so it is not trapped into a local minimum in the estimation of the C_f parameter. The cost function of this optimization process is the sum of the squares of the residuals. The method returns the size $C_f^{'}$ of the non-linear cascade, and the slope $a_f^{'}$ of the sigmoidal curve minimizing the cost function. After the estimation of the f^{th} non-linear part of the model, the flag NL_{in} is set to false (line *68*), so it can be used in the estimation process of a possible subsequent non-linear adoption phase. Also, the flag $NLmdl_{est}$ is set to true in line *69*, since a new non-linear model has just been estimated thereby allowing the generation of forecasts.

Construction of the mathematical expression of the model. If the process has already added linear terms ($k > 0$) to the model, its mathematical expression after the addition of the f^{th} non-linear term is given by the formula in line *71*. If the process has not encountered any linear growth phase, then the model contains only the f non-linear terms and an intercept value (line *73*). In this case the adoption process starts with a non-linear phase and continues with a succession of such phases, whereby each one superimposes a sigmoidal growth pattern onto the end point of the previous one.

Generation of forecasts during non-linear growth phases. In line *75* the process produces a forecast $\widehat{N(t_{nl(f,2)})}$ of the cumulative number of adopters at the end of the f^{th} non-linear phase. This number is equal to the total number of adopters $N(t_{nl(f,\,1)})$ at the start point of the f^{th} non-linear phase plus the estimated number of adopters at the end of this non-linear phase. This value is returned in line *76*. When the process is within a non-linear adoption phase, whose parameters have already been estimated (line *78*), forecasts about the total number of adopters at time points before the expected end of the current non-linear phase are generated according to the formula described in line *80*. Prior to the generation of a forecast the user defines a time point less than or equal to the time point at which the non-linear phase is expected to end (line *79*). The forecast is returned to the user in line *81*. When the condition of the *"if"* statement in line *78* is not satisfied because the current non-linear phase has ended (*"else-if"* in line *83*), the process sets the flag NL_{ex} to *"true"* (line *84*), so it can force a re-estimation of the linear growth trend after the end of the non-linear phase. This happens when the acceleration of the adoption rate is lower than Δ, as described in lines *11* and *12*. The aforementioned process is repeated until the final time period T.

6.3. Time complexity of the prediction method algorithm

The proposed algorithm aims to generate forecasts in real time during a period of T time intervals of predefined length. The *"for loop"* in line 6 is not considered in the calculation of the complexity of the algorithm, since it only deals with the repetition of the prediction process during the monitored period. At each time point t_i the input to the algorithm is just a single integer representing the number of new adopters at t_i. With the exception of the parameter estimation task, all the other tasks of the algorithm are executed in constant time, since they consist of elementary operations such as comparisons, assignment of values to flags, filling and accessing array elements, and arithmetic calculations. The time complexity of these operations is $O(1)$. The *"if-then-else"* parts of the algorithm consist of basic operations, and therefore their time complexity is also $O(1)$.

The estimation of the parameters of the predictive model is carried out through the least squares (linear/non-linear) fitting method whose time complexity is $O(MP^2)$, where M is the number of data points used for the estimation of the model, and P is the number of the estimated parameters. In the experiments of this study the estimation of the stationary adoption rate was carried out using 4-5 data points, which were enough for the generation of accurate forecasts. As a result, we can consider the time complexity of this task as constant, since the number of data points and estimated parameters are fixed. To account for the case of more data points being required for the estimation of the stationary trend, we treat this number as a user defined parameter dp_s. This means that the time complexity of the estimation of a 1^{st}-degree polynomial model capturing the popularity growth dynamics during stationary periods is $O(dp_s * 1^2) = O(dp_s)$, since we estimate one parameter.

For the estimation of the two parameters of a non-linear part of the model, that is the size C_f of a cascade and the slope of the sigmoidal curve a_f, we need the data points from the lower asymptote of the sigmoidal popularity growth curve to its inflection point. While this number varies among non-linear phases, the short-lived nature of non-linear periods implies that this number is actually small. Let M denote the number of data points for the estimation of the two parameters (C_f, a_f) of a non-linear adoption phase, then the time complexity of this parameter estimation task is $O(M * 2^2) = O(4M)$, which is equivalent to $O(M)$. Since this is the dominant term we can infer that the worst case time complexity of the proposed algorithm is $T(n) = O(M)$. The best case time complexity, that is, the complexity when there is no need to estimate a non-linear part of the model is $O(1)$ (constant time), given that the number of data points for the estimation of a stationary adoption trend is fixed.

7. Discussion and conclusion

To predict the popularity of online content we followed a three-stage approach comprising the analysis of empirical data, the construction of a model describing the identified popularity evolution patterns, and finally the use of this model in prediction tasks. In particular, we modeled the popularity growth of online content as a sequence of linear and non-linear phases stemming from stationary and non-stationary adoption periods characterizing the propagation of any type of online content. In most of the experiments of this study we identified a single, stable, linear growth trend, on which sigmoidal growth patterns were superimposed each time an intervention event affected the popularity evolution of the examined online content. We formulated a general equation-based model incorporating the dynamics of multiple linear and non-linear growth periods, and a simpler version featuring a single linear and various non-linear phases. We showed that these models accurately describe the empirical popularity patterns of persistent and non-persistent, popular and non-popular online content, since they capture the dynamically changing statistical properties of the popularity evolution. As such they are capable of generating popularity forecasts of high precision through extrapolation, thus substantially improving on the effectiveness of existing methods. By organizing the steps of the proposed prediction method into an algorithm, we provided a technical-oriented description addressing the implementation of a process, suitable for the real time prediction of the popularity growth of online content without training, using minimal input information.

In regards to the statistical characteristics of the non-linear popularity growth patterns, we observed that they are invariant across different timescales. Specifically, the sigmoidal growth phases present a self-similar, fractal-like structure, with the non-linear adoption cascades observed at larger timescales being made up of qualitatively similar cascades occurring at smaller divisions of time. Such a property makes the proposed methodology applicable and equally reliable to prediction tasks of microscopic, mesoscopic, and macroscopic timescales. The ability of the proposed method to generate forecasts at small timescales is particularly important for the popularity prediction of fleeting online content, since the popularity evolution of this type of content is ephemeral, thereby necessitating short-term predictions. The analysis that we performed implies that the overall popularity growth pattern depends on the density and intensity of perturbations caused by intervention events, which superimpose non-linear popularity growth phases on a linear adoption trend characterizing the online content propagation for significant lengths of time. The extent to which an initial perturbation can trigger the appearance of subsequent perturbations in an autocatalytic way, determines the potential of a piece of online content to be adopted by discrete subpopulations of higher threshold individuals. The popularity growth pattern generated by intervention events can be described by a sigmoidal function, whose saturation level and slope correspond to the number of opportunistic adopters and the level of their tuning to an exogenous stimulus. Furthermore, the gradient of the sigmoidal adoption curve provides information about the distribution of the adoption thresholds of the users who responded to an intervention event. Denoting the slope of a sigmoidal curve by a, the quantity $1/a$ is directly related to the variance of the adoption thresholds.

The proposed popularity prediction method not only provides popularity forecasts, but also answers questions concerning the dynamical state of an online content diffusion process. For instance, by measuring and monitoring the acceleration of the adoption rate, we can infer the evolution pattern of a non-linear adoption cascade. In particular, while the accelerometer shows a positive increasing trend, it can be inferred that the adoption cascade size will expand to at least twice its current size. When the acceleration becomes zero, meaning that the sigmoidal adoption growth curve has reached its inflection point, it can be easily guessed that the cascade size will approximately double in a time period roughly equal to the period from the beginning of the sigmoidal curve to its inflection point. When the acceleration of the adoption rate acquires a negative value it can be safely inferred that the evolving cascade approaches its final size. Another important prediction question answered by the proposed method is the early forecast of the size of a non-linear adoption cascade, when we know in advance the inflection point of the sigmoidal popularity growth. Such situations are quite common, thereby making the proposed method a highly useful tool for generating forecasts in various socio-economic contexts.

While the size of the non-linear adoption cascades is critical to a successful dissemination of online content through intervention events, the stationary adoption trend is equally important. Actually, in the long run it is through this propagation mode that the online content reaches different subpopulations, in a manner similar to signal transmission in a homogeneous medium. Subsequently, the newly exposed users might collectively respond to other intervention events, thereby generating new non-linear adoption cascades. The study of the network topological properties, and of the behavioral characteristics of the content adopters during the linear and non-linear growth periods, can provide significant knowledge concerning their heterogeneity level, and how an online social network interacts with its surroundings. Such insights can enable the identification of information channels of high conductance, thus allowing the study of their efficiency in the transmission of online content and of their reaction to exogenous stimuli. Furthermore, it is particularly interesting to investigate the interface between online social networks and their external environment, with a view to detecting how the probabilistic properties of a content diffusion process are affected by dynamical processes occurring in the surroundings of social networks. Also, the process through which an intervention event causes subsequent perturbations in an autocatalytic way should be explored, in order to extend the understanding of the mechanisms underlying the appearance of non-linear cascades, thereby substantially contributing to their early prediction. Another important field of further investigation relates to the statistical properties of stationary diffusion processes. The examination of the factors determining the average adoption rate, which stays almost constant during stationary periods, can yield insights into potential ways which can increase the mean diffusion speed, thus leading to a self-sustained, enduring adoption trend requiring minimal external intervention for a widespread diffusion of online content.

The proposed popularity prediction method is based on the statistical properties of the adoption process of online content. It was demonstrated that these properties are universal to problems with different timescales and different types of content, thereby making the new method applicable to any online context. The simplicity, accuracy, robustness, and universality of the proposed method make it a useful, practicable and reliable tool for predicting the popularity of online content using minimal, real time, publicly available information.

Acknowledgements

The Author wishes to extend his gratitude to the Editor-in-Chief, Professor Witold Pedrycz, and the anonymous Associate Editor and three Reviewers for their detailed comments and suggestions that greatly improved the manuscript. Also, the Author is deeply grateful to Professor Steven Bishop, Department of Mathematics, University College London, for providing help in the language editing of the manuscript, for reviewing its content, and for making suggestions for improving its comprehensibility.

References

[1] M. Ahmed, S. Spagna, F. Huici, S. Niccolini, A peek into the future: predicting the evolution of popularity in user generated content, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, ACM, 2013, pp. 607–616.
[2] S. Asur, B.A. Huberman, G. Szabo, C. Wang, Trends in social media: persistence and decay, in: Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM), 2011.
[3] R. Bandari, S. Asur, B.A. Huberman, The pulse of news in social media: forecasting popularity., in: Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM), 2012, pp. 26–33.
[4] P. Bao, H.-W. Shen, J. Huang, X.-Q. Cheng, Popularity prediction in microblogging network: a case study on sina weibo, in: Proceedings of the 22nd International Conference on World Wide Web companion, International World Wide Web Conferences Steering Committee, 2013, pp. 177–178.
[5] F.M. Bass, A new product growth for model consumer durables, Manage. Sci. 15 (5) (1969) 215–227.
[6] R.G. Box G., G. Jenkins, Time Series Analysis: Forecasting & Control, fourth ed., Wiley, 2008.
[7] M. Cha, A. Mislove, K.P. Gummadi, A measurement-driven analysis of information propagation in the flickr social network, in: Proceedings of the 18th International Conference on World Wide Web, ACM, 2009, pp. 721–730.
[8] J. Cheng, L. Adamic, P.A. Dow, J.M. Kleinberg, J. Leskovec, Can cascades be predicted? in: Proceedings of the 23rd International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2014, pp. 925–936.
[9] D.R. Cox, Renewal Theory, vol. 4, Methuen London, 1962.
[10] R. Crane, D. Sornette, Robust dynamic classes revealed by measuring the response function of a social system, Proc. Nat. Acad. Sci. 105 (41) (2008) 15649–15653.
[11] P. Cui, S. Jin, L. Yu, F. Wang, W. Zhu, S. Yang, Cascading outbreak prediction in networks: a data-driven approach, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2013, pp. 901–909.
[12] P.S. Dodds, D.J. Watts, A generalized model of social and biological contagion, J. Theor. Biol. 232 (4) (2005) 587–604.
[13] F. Figueiredo, On the prediction of popularity of trends and hits for user generated videos, in: Proceedings of the sixth ACM International Conference on Web Search and Data Mining, ACM, 2013, pp. 741–746.

[14] S. Gao, J. Ma, Z. Chen, Modeling and predicting retweeting dynamics on microblogging platforms, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, ACM, 2015, pp. 107–116.
[15] J.W. Gibbs, Elementary Principles in Statistical Mechanics, Courier Corporation, 2014.
[16] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, Y. Moreno, The dynamics of protest recruitment through an online network, Scientif. Rep. 1 (2011).
[17] M. Granovetter, Threshold models of collective behavior, Am. J. Sociol. (1978) 1420–1443.
[18] M. Gupta, J. Gao, C. Zhai, J. Han, Predicting future popularity trend of events in microblogging platforms, Proc. Am. Soc. Inf. Sci. Technol. 49 (1) (2012) 1–10.
[19] G. Gursun, M. Crovella, I. Matta, Describing and forecasting video access patterns, in: INFOCOM, 2011 Proceedings IEEE, IEEE, 2011, pp. 16–20.
[20] M. Jenders, G. Kasneci, F. Naumann, Analyzing and predicting viral tweets, in: Proceedings of the 22nd International Conference on World Wide Web companion, International World Wide Web Conferences Steering Committee, 2013, pp. 657–664.
[21] S. Kong, Q. Mei, L. Feng, F. Ye, Z. Zhao, Predicting bursts and popularity of hashtags in real-time, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, 2014, pp. 927–930.
[22] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, A. Kustarev, Prediction of retweet cascade size over time, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ACM, 2012, pp. 2335–2338.
[23] J.G. Lee, S. Moon, K. Salamatian, Modeling and predicting the popularity of online contents with cox proportional hazard regression model, Neurocomputing 76 (1) (2012) 134–145.
[24] K. Lerman, T. Hogg, Using a model of social dynamics to predict popularity of news, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 621–630.
[25] J. Leskovec, L. Backstrom, J. Kleinberg, Meme-tracking and the dynamics of the news cycle, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009, pp. 497–506.
[26] H. Li, X. Ma, F. Wang, J. Liu, K. Xu, On popularity prediction of videos shared in online social networks, in: Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management, ACM, 2013, pp. 169–178.
[27] I.N. Lymperopoulos, G.D. Ioannou, Online social contagion modeling through the dynamics of integrate-and-fire neurons, Inf. Sci. 320 (0) (2015) 26–61.
[28] Z. Ma, A. Sun, G. Cong, On predicting the popularity of newly emerging hashtags in twitter, J. Am. Soc. Inf. Sci. Technol. 64 (7) (2013) 1399–1410.
[29] Y. Matsubara, Y. Sakurai, B.A. Prakash, L. Li, C. Faloutsos, Rise and fall patterns of information diffusion: model and implications, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 6–14.
[30] J.D. Murray, Mathematical Biology I: An Introduction, Interdisciplinary Applied mathematics, vol. 17, 2002.
[31] N. Naveed, T. Gottron, J. Kunegis, A.C. Alhadi, Bad news travel fast: A content-based analysis of interestingness on twitter, in: Proceedings of the 3rd International Web Science Conference, ACM, 2011, p. 8.
[32] S. Petrovic, M. Osborne, V. Lavrenko, Rt to win! predicting message propagation in twitter., in: Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM), 2011.
[33] H. Pinto, J.M. Almeida, M.A. Gonçalves, Using early view patterns to predict the popularity of youtube videos, in: Proceedings of the sixth ACM International Conference on Web Search and Data Mining, ACM, 2013, pp. 365–374.
[34] E.M. Rogers, Diffusion of Innovations, Simon and Schuster, 2010.
[35] H.-W. Shen, D. Wang, C. Song, A.-L. Barabási, Modeling and predicting popularity dynamics via reinforced poisson processes, arXiv preprint arXiv: 1401.0778(2014).
[36] G. Szabo, B.A. Huberman, Predicting the popularity of online content, Commun. ACM 53 (8) (2010) 80–88.
[37] A. Tatar, M.D. de Amorim, S. Fdida, P. Antoniadis, A survey on predicting the popularity of web content, J. Internet Serv. Appl. 5 (1) (2014) 1–20.
[38] M. Tsagkias, W. Weerkamp, M. De Rijke, Predicting the volume of comments on online news stories, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM, 2009, pp. 1765–1768.
[39] O. Tsur, A. Rappoport, What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities, in: Proceedings of the fifth ACM International Conference on Web Search and Data mining, ACM, 2012, pp. 643–652.
[40] T.W. Valente, Social network thresholds in the diffusion of innovations, Soc. Netw. 18 (1) (1996) 69–89.
[41] M. Vasconcelos, J.M. Almeida, M.A. Gonçalves, Predicting the popularity of micro-reviews: a foursquare case study, Inf. Sci. 325 (2015) 355–374.
[42] P.-F. Verhulst, Notice sur la loi que la population suit dans son accroissement. correspondance mathématique et physique publiée par a, Quetelet 10 (1838) 113–121.
[43] L. Weng, F. Menczer, Y.-Y. Ahn, Virality prediction and community structure in social networks, Scientif. Rep. 3 (2013).
[44] T. Wu, M. Timmers, D.D. Vleeschauwer, W.V. Leekwijck, On the use of reservoir computing in popularity prediction, in: Evolving Internet (INTERNET), 2010 Second International Conference on, IEEE, 2010, pp. 19–24.
[45] J. Yang, J. Leskovec, Patterns of temporal variation in online media, in: Proceedings of the fourth ACM International Conference on Web Search and Data Mining, ACM, 2011, pp. 177–186.
[46] T. Zaman, E.B. Fox, E.T. Bradlow, et al., A bayesian approach for predicting the popularity of tweets, Ann. Appl. Stat. 8 (3) (2014) 1583–1611.
[47] T.R. Zaman, R. Herbrich, J. Van Gael, D. Stern, Predicting information spreading in twitter, in: Workshop on computational social science and the wisdom of crowds, nips, 104, Citeseer, 2010, pp. 17599–17601.
[48] P. Zhang, X. Wang, B. Li, On predicting twitter trend: factors and models, in: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, 2013, pp. 1427–1429.
[49] Q. Zhao, M.A. Erdogdu, H.Y. He, A. Rajaraman, J. Leskovec, Seismic: A self-exciting point process model for predicting tweet popularity, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 1513–1522.

Ilias N. Lymperopoulos is a Computer Engineer who graduated from the Department of Computer Engineering and Informatics at the University of Patras in Greece. In 2011 he obtained his MBA diploma with the highest degree excellence (ranked 1st) from the Athens University of Economics and Business. From April 2012 to September 2015, he carried out doctoral studies at the Department of Management Science and Technology of the Athens University of Economics and Business, where he continues conducting research as postdoctoral fellow. His research interests focus on the study and modeling of online social dynamical processes pertaining to the emergence of collective phenomena, formation of online activity patterns, social transmission, and information diffusion. Inspired by biological neural networks, he investigates social dynamical process through the lens of neuroscience, artificial intelligence and complex adaptive systems. He has thorough managerial and technical professional experience acquired from his participation with leading roles in large and strategic IT projects in various fields such as Robotics, ERP systems, eCommerce, Telecommunications, Transport Telematics, Logistics and Banking.