On Popularity Prediction of Videos Shared in Online Social Networks

Haitao Li
Simon Fraser University
Burnaby, BC, Canada
haitaol@sfu.ca

Xiaoqiang Ma
Simon Fraser University
Burnaby, BC, Canada
xma10@sfu.ca

Feng Wang
The University of Mississippi
University, MS, USA
fwang@cs.olemiss.edu

Jiangchuan Liu
Simon Fraser University
Burnaby, BC, Canada
jcliu@cs.sfu.ca

Ke Xu
Tsinghua University
Beijing, China
xuke@mail.tsinghua.edu.cn

ABSTRACT

Popularity prediction, with both technological and economic importance, has been extensively studied for conventional video sharing sites (VSSes), where the videos are mainly found via searching, browsing, or related links. Recent statistics however suggest that online social network (OSN) users regularly share video contents from VSSes, which has contributed to a significant portion of the accesses; yet the popularity prediction in this new context remains largely unexplored. In this paper, we present an initial study on the popularity prediction of videos propagated in OSNs along friendship links.

We conduct a large-scale measurement and analysis of viewing patterns of videos shared in one of largest OSNs in China, and examine the performance of typical views-based prediction models. We find that they are generally ineffective, if not totally fail, especially when predicting the early peaks and later bursts of accesses, which are common during video propagations in OSNs. To overcome these limits, we track the propagation process of videos shared in a Facebook-like OSN in China, and analyze the user viewing and sharing behaviors. We accordingly develop a novel propagation-based video popularity prediction solution, namely SoVP. Instead of relying solely on the early views for prediction, SoVP considers both the intrinsic attractiveness of a video and the influence from the underlying propagation structure. The effectiveness of SoVP, particularly for predicting the peaks and bursts, have been validated through our trace-driven experiments.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Sociology; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505523.

Keywords

Social network, video sharing, popularity prediction, propagation

1. INTRODUCTION

In the past decade, online social networks (OSNs) (e.g., Facebook, Twitter, Google+, and etc.) have become popular online destinations for connecting friends as well as sharing contents. Traditionally, a user finds videos by browsing the front pages or related video lists in such video sharing sites (VSSes) as YouTube, or via search engines [38]. The emergence of OSNs however has greatly changed such access patterns through proactively and efficiently sharing among friends the video links from external VSSes [24]. The latest statistics by YouTube indicate that 500 years of YouTube video are watched every day by Facebook users, and over 700 YouTube videos are shared on Twitter each minute nowadays [36]. The comScore's statistics [6] in August 2012 further reveal that Facebook has ranked eighth in terms of video content views. Besides Facebook and Twitter, we have seen similar trend around the world. For example, as of May 2011, more than 54 million unique RenRen [1] (the largest Facebook-like OSN in China) users have participated in video viewing and 20 million participated in sharing, generating 12.4 million views, and 1.64 million shares every day [17].

Content providers, advertisers, and Web hosts all expect to predict how many view accesses the individual videos might generate to a given site. For advertising, the popularity count is tied directly with the ad revenue (see for example the ads shown with YouTube videos); an accurate population prediction thus offers a good revenue (or cost) indication for both YouTube and its content generators. For content-distribution networks, the computation, storage, and bandwidth resources can be well planned with a good prediction of the access patterns [31, 18]. There have been extensive studies on popularity prediction for conventional VSSes, mostly leveraging earlier views of a video as the key predictor [30, 21, 9, 26, 34].

Although the videos shared in OSNs are generally hosted by VSSes, an OSN proactively spreads videos among its users along friendship relations. As such, a video's views are

[1] www.renren.com

not only determined by the users' interest in it, but also the underlying propagation structure, which generates unique request patterns than that in VSSes. It has been found that the propagation-based video spreading mechanism generates distinguished video popularity distribution [17]. We further find that it would lead to high video popularity dynamics due to great difference of the numbers of users' friends. As such, even though it is proved that the conventional prediction models perform well in predicting video views in VSSes [30], it is necessary to evaluate their effectiveness in the OSN context and if needed, to develop new tools.

In this paper, we conduct an initial study on the popularity prediction of videos shared in OSNs. Collaborated with a large Facebook-like OSN in China, we first measure and analyze the characteristics of video popularity evolutions in this large OSN. We then test the performance of conventional views-based prediction models, and also propose an novel propagation-based prediction solution. Our contributions are summarized as follows:

- By analyzing long-term traces of video views, we find that video popularity evolution in the OSN is highly dynamic, where the correlations between the views in early and later times are noticeably lower than that in VSSes. The lower correlations pose significant challenge to views based prediction tools.
- We test the performances of the conventional prediction tools including Autoregressive Integrated Moving Average (ARIMA) model, Multiple Linear Regression (MLR), and k-Nearest Neighbors (kNN). These models only need the number of early views as the input, and can be easily developed by VSSes without assistances of OSNs. We find that they are generally ineffective, if not totally fail, especially when predicting the early peaks and later bursts of accesses, which are common during video propagations in OSNs.
- We present a novel propagation-based prediction tool, namely SoVP (Social network assisted Video Prediction). SoVP considers both the intrinsic attractiveness of a video and the influence from the underlying propagation structure. The effectiveness of SoVP, particularly for predicting the request bursts, has been validated through our trace-driven experiments.

The rest of the paper is organized as follows. We introduce some related work in Section 2. Section 3 introduces measurement methodology and depicts the characteristics of video popularity evolution in the OSN. Section 4 introduce the premier knowledge of three conventional views-based prediction models. We propose a novel propagation-based prediction framework in Section 5. Section 6 presents trace-based evaluations. We conclude in Section 7.

2. RELATED WORK

Popularity prediction of online content has been widely studied in the literature. Earlier studies have focused on predicting the spread of information based on time series. Typical solutions include *time series* models like ARIMA [21, 9], *regression* models [32, 13, 25, 30, 34, 35], and *classification* models [32, 26, 27]. For video prediction, they predicted the future views solely based on the early views, which we refer to as *views-based* predictions. Their efficiency highly depends on the characteristic of the data set. Cha *et al.* [2] found that, in YouTube, a high linear correlation existed between the number of video views on early days and later days (e.g., correlation coefficient is 0.84 between the 2^{nd} day and the 90^{th} day). Szabo *et al.* [30] also found similar results and presented three models using linear correlation and regression for prediction. These models can predict video popularity 30 days ahead with a remarkable accuracy (e.g., relative error of 10%) based on 10-day historic video views. Pinto *et al.* [22] proposed two models for predicting the future popularity of the YouTube video by learning its early view patterns. In this paper, we study the video accesses through OSN sharing, which is quite different from the conventional YouTube-like accesses [17, 16]; we have examined whether the above conventional models can well predict popularity in this new context and the results are largely negative.

Recently there have been pioneering data-driven analysis of information propagation in different kinds of OSNs, e.g., photos propagation in Flick network [3], likes and fans pages in Facebook [1, 29, 33], links and retweets in Twitter [11, 24, 4, 8, 14, 19, 37], and voting in Digg [14, 28, 15]. There have also been efforts towards prediction in this context [8, 11, 15]. Galuba *et al.* [8] proposed a propagation model that predicts which users are likely to mention which URLs in Twitter. Hong *et al.* [11] treated the retweets prediction in Twitter as a classification task. They investigated a wide spectrum of features to determine which ones can be successfully used as predictors of popularity. Kooti *et al.* [12] investigated the prediction of emerging social conventions on Twitter. The most close research to ours was conducted by Lerman *et al.* [15]. They predicted popularity of news in Digg, by incorporating aspects of the web site design. They showed that their model-based prediction improves prediction based on simply extrapolating from the early votes. Our work has been inspired by these studies, and differs from theirs in that we focus on video, which, as one of the most information-rich data objects, preserves unique characteristics that are yet to be examined for prediction.

3. VIDEO PROPAGATION AND POPULARITY EVOLUTION

This section introduces our measurement methodology, and depicts the characteristics of video propagation and popularity evolution in the OSN.

3.1 Measurement Methodology

To understand video spreading in OSNs, we closely collaborate with a large-scale Facebook-like OSN in China to collect and analyze its video-related user behaviors [2]. Like Facebook, its users primarily interact with information through an aggregated history of their friends' recent activity, called the "News Feed". For video sharing, typically a user may post a video link from a VSS, and the link will appear in its friends' "News Feed". Some friends may click and view the video, and such viewers can then decide whether to reshare the video. If they click the "share" button, the video link will appear in their friends' "News Feed" and hence the video can further propagate.

The data collection process works as follows: when a user clicks a video link shared by her/his friend, a record will be

[2]To protect user privacy, we translate real UserIDs by some hash function, and user IPs are not included in our date set.

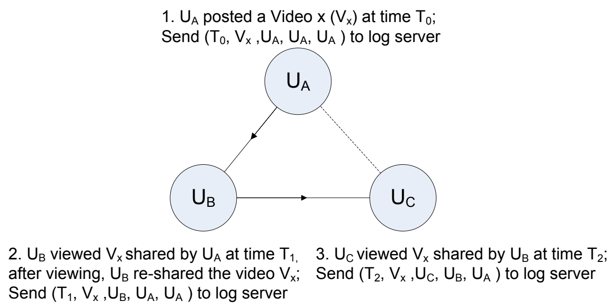

Figure 1: Illustration of video propagation and corresponding logs

sent to a log server; and the data format is: (*Starting Time*, *Video URL*, *Viewer ID*, *Direct Sharer ID*, *Initial Sharer ID*). We use an example in Fig. 1 to illustrate the video propagation and the corresponding log record. Initially at time T_0, user A (denoted as U_A) posted Video x (denoted as V_x) from a VSS, and then a record (T_0, V_x, U_A, U_A, U_A) is sent to log server. Since U_A is the initial user, both direct sharer and initial sharer are itself; At time T_1, U_B viewed V_x through the share link created by U_A, and then U_B further shared V_x after watching it; and then a record (T_1, V_x, U_B, U_A, U_A) is sent to log server. Also as U_A is the initial user, the initial sharer is U_A; At the Time T_2, U_C viewed V_x through the share link created by U_B. A new record (T_1, V_x, U_C, U_B, U_A) is sent to log server. Note that there is a dotted line without any arrow between the friends U_A and U_C, which means although U_A's shared video was exposed in U_C's "News Feed", U_C did not click it maybe because s/he is offline.

Table 1: Summary of trace in one-day period

Views	Shares	Users	Videos	New Videos
12,432,708	1,628,852	3,514,461	201,517	71,236

Using (Video URL, Viewer ID), we can extract the number of views of any video in each day. We then use this information to analyze the video popularity evolution patterns, and test views-based prediction models. Using (Video URL, Viewer ID, Direct Sharer ID), we can examine the share-view relationship between two friends. And together with the initial Sharer ID, we can restore a video's propagation process. Such information is useful to analyze the reason underlying the popularity evolution patterns, and inspire the design of our propagation-based prediction model. Our study in this paper is based on a one-month trace that began from March 24^{th}, 2011, since we find that most requests of a video are generally cumulated in the first month, and after that the daily requests decline to a very small scale. Table 1 presents the statistics in a typical one-day period (March 24^{th}, 2011) during the measurement. Our records covered all video requests in the measurement period. In the one-month period, we recorded about 370 million views and 49 million shares.

3.2 Video Propagation

A common video propagation process is like this: Initially, a user shares a video link to an OSN directly from a VSS. Immediately, this user's friends can find this video in their "News Feed", and some of them watch this video. After that, some portion of these viewers will share this video and can recommend it to their friends. To specify this process, we give the following definitions. We call the users in the root of a propagation tree $initiators$. These users are the ones who independently shared the video directly from VSSes. We call the users who re-shared the video $spreaders$. We call the users who watched the shared video $viewers$. Since spreaders generally watched the video before re-shared it, most of them are also viewers. The definition of $viewers$ is different from that in [12, 20]. In their model, the $viewers$ are exclusive of spreaders. We define a video's $popularity$ as the number of its $viewers$. We define the $BranchingFactor(BrF)$ as the number of $viewers$ directly follow a $spreader$. We define the $ShareRate(ShR)$ as the ratio of the $viewers$ that re-share the video after watching it.

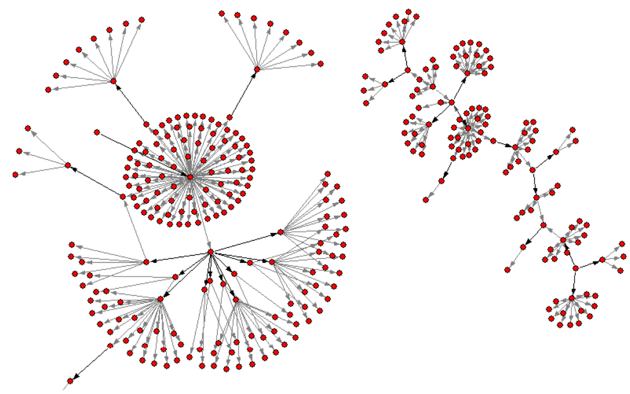

Figure 2: Illustration of a video propagation

The video propagation of popular videos are very complex. For example, we find one video which consists of 1022 initiators, 153185 spreaders, and 995707 viewers over one month propagation. Each of 1022 propagation trees exhibits unique patterns. We choose two among them and illustrate their propagation structures over several hours in Fig. 2. Each vertex is a user and the arrows means that a user has viewed the video shared by his/her friend. We can observe some super spreaders in the left tree, who are followed by hundreds of viewers, while the spreaders in the right tree attract moderate viewers. The two different trees from the same video gives us an illustration that the underlying OSN topology plays a foundational role in video propagation and popularity evolution.

3.3 Popularity Evolutions of Typical Videos

According to a video's attractiveness (ShR and BrF), we roughly classify popular videos into three types [3]: high BrF & high ShR, high BrF & low ShR, and low BrF & high ShR. Although finer classifications like the work in [7] would be possible and worth further study, current classification is enough to explore the limits of conventional models in predicting popularity of videos shared in OSNs.

We choose one typical video from each type and show them in Fig. 3, 4 and 5, respectively. The middle and lower sub-figures show the evolution of ShR and BrF. The upper

[3]Since the paper concentrates on popular videos, the category low BrF & low ShR is not mentioned, which generally refers to unpopular videos (e.g., less than 10 views per day).

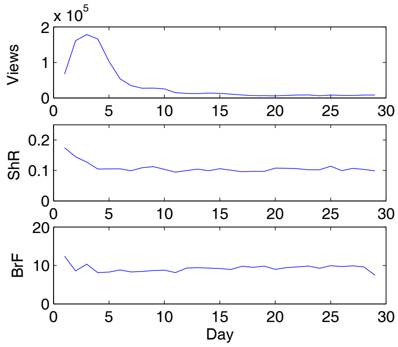

Figure 3: Popularity evolution of the type-1 video (high BrF, high ShR)

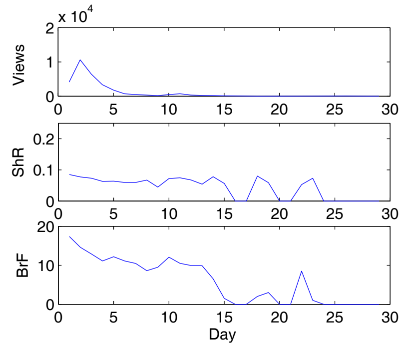

Figure 4: Popularity evolution of the type-2 video (high BrF, low ShR)

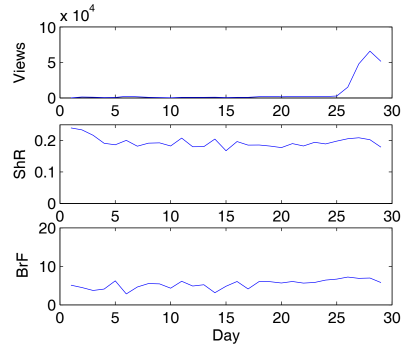

Figure 5: Popularity evolution of the type-3 video (low BrF, high ShR)

sub-figures show the evolution of video views in each day. The type-1 video was the most popular video in our sample videos. It kept the views at a very high level during the first week. Although experiencing decreasing views after that, it still received more than seven thousands views after one month. Like the type-1 video, the type-2 video also experienced a surge-growth over first few days (e.g., two days), acquiring huge (e.g., 90%) views. Yet different from the type-1 video, it quickly turned to the sluggish state after the peak, only receiving less than a hundred of views every day after one week. The type-3 video stayed dormant for several days (e.g., nearly one week) after they were first shared in the OSN; then it experienced a dramatic increase and attracted a large portion of total views within a few days. Overall, while the video shared in OSNs generally experiences a request burst, it is uncertain about the start time, the height and duration of the burst. In the performance evaluation section we will find these uncertainties pose challenges to conventional views-based prediction models.

3.4 Correlation between Early and Later Views

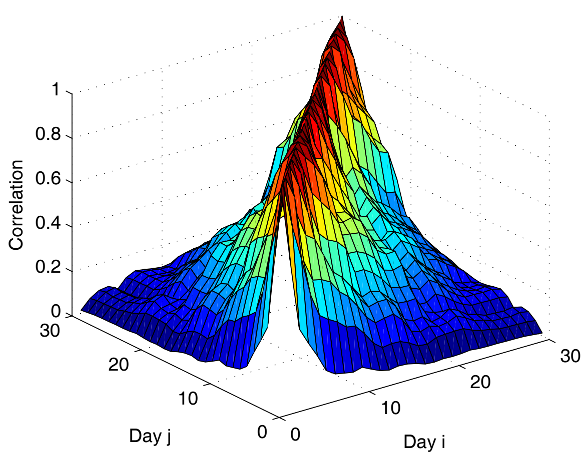

Figure 6: Correlation between early and later views

Similar to the previous works [2, 30], we examine the correlation between early and later views, which is a simple but effective indicator to show whether the number of early views is an effective factor for the prediction of future views. For the span of 30 days, we compute the Pearson correlation coefficients [23] in terms of the number of views across the top-2% videos at early and later days and show the result in Fig. 6. Both early day and later day vary from 1 to 30. We can see that the correlation is very high when the later day is within 2-3 days of the early day, and becomes very small when the later day is out of this range. This contradicts the conclusion in the previous works that the correlation is still very high even when the later day is tens of days after the early day. As such, we are interested in whether conventional views-based prediction models still work well, and thus we conduct a comprehensive comparison study, as discussed in the following.

4. VIEWS-BASED PREDICTION

One target of this paper is to investigate whether the number of future (e.g., one-day ahead) views can be accurately predicted simply based on early views, which can be easily obtained by VSSes so that they can do predictions without assistances of OSNs. To do this, we will examine three conventional prediction models: ARIMA [21], MLR [25], and kNN [20]. To make predictions, they either utilize the early views of the predicted video itself or utilize the similarity of the popularity evolution pattern with early published videos. Here we provide some primary knowledge of these models, and present their performance in Section 6.

4.1 Autoregressive Integrated Moving Average (ARIMA)

We first examine Autoregressive Integrated Moving Average (ARIMA), one of the most popular time series models for predicting future values of a time series [21, 9]. Given the time series of video popularity in the past several days, it can make fine-grained prediction for the video's future evolution, leveraging the trend, periodicity and autocorrelation exhibited in the history information. ARIMA consists of three parts: an Autoregressive (AR) model, a Moving Average (MA) model and an integrated part. They are applied in the cases where data show evidence of non-stationarity and an initial differencing step (corresponding to the "integrated" part of the model) can be used to remove the non-stationarity. Given a time series Y, an AR model of order p is defined as:

$$Y(t) = \sum_{i=1}^{p} \beta_i Y(t-i) + \epsilon \tag{1}$$

where $Y(t)$ is the number of views in the t^{th} day; β_1,..., β_p are the parameters of the model; and ϵ is a white noise error term. An MA model of order q is defined as follows:

$$Y(t) = \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{2}$$

where θ_1,...,θ_q are the parameters of the model and ϵ_t,...,ϵ_1 are again white noise error terms. Combing Eq. 1 and 2, an ARIMA model of order (p, q) is written as follows:

$$Y(t) = \sum_{i=1}^{p} \beta_i Y(t-i) + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{3}$$

The error terms, ϵ_t, are generally assumed to be Gaussian random variables with zero mean and constant variance.

4.2 Multiple Linear Regression (MLR)

A major drawback of ARIMA model is that it needs a relatively long period of history information for prediction. For our data set, the number of views of at least first 4 days are required to generate the model and thus the initial population evolution for a newly released video cannot be predicted using ARIMA. The high correlation of neighbor days motivates us to use regression models. Multiple Linear Regression (MLR) [25] is widely used to model the relationship between a dependent variable and several explanatory variables. In our scenario, early views are regarded as explanatory variables and used to predict later views, which is shown in Eq. 4:

$$Y(t) = \alpha + \sum_{i=t-n}^{t-1} \beta_i Y(i) + \epsilon_t \tag{4}$$

where $Y(t)$ is the number of views in the t^{th} day; α is a constant number; β_i is the weight for the i^{th} day; and ϵ_t is the residual value. n is the critical parameter in this model that defines the number of early days used for prediction.

4.3 k-Nearest Neighbors Regression (kNN)

kNN regression [20] is also a widely used regression model. It estimates the value of an unknown function at a given point based on the values of its nearest neighbor points. The kNN estimator is defined as the weighted average function value of the nearest neighbors. In our scenario, the views of the videos in the training set are used to predict the views of the videos in the test set, as shown in Eq. 5:

$$Y_x(t) = \sum_{x' \in N(x)} \frac{1/d(x,x')}{\sum_{x'' \in N(x)} 1/d(x,x'')} Y_{x'}(t) \tag{5}$$

where $Y_x(t)$ is the number of views of video x in the t^{th} day; $N(x)$ is the set of k nearest points to video x in the training set with regard to the views in previous days; $d(\cdot)$ denotes the distance function; and k is the parameter defining the number of neighbors. We choose Euclidean distance as the distance function. Similar to MLR, we use the early views as the vector to compute the distance between future days. To break ties in neighbor selection, we include all the videos with equal distance since the late views can vary a lot with equal early views, especially when only a short period of early views are considered.

5. PROPAGATION-BASED PREDICTION

Comparing with VSSes, OSNs know much more information about a video beyond the number of its early views, such as viewers, sharers, whether viewers would like to share the video after viewing, whether users would like to view the videos shared by their friends, and etc.. Yet, how to utilize such information in video popularity prediction is not easy, as the previous work has shown that they have no simple (e.g., linear) relationship with the video popularity [16]. In this section, we propose a novel propagation-based prediction framework to predict video future views in the OSN.

5.1 Modeling Video Propagation

Before modeling the video propagation, we first define some notations. For a given video, $V(t)$ and $S(t)$ are defined as the sets of its viewers and sharers by the time t, respectively. We use $|V(t)|$ to denote the number in the set $V(t)$, and this notation can also apply to other sets such as $S(t)$. $ShR(t)$ (short for *Sh*aring *R*ate) is the probability that a user will reshare a video after viewing it. $ViR(t)$ (short for *Vi*ewing *R*ate) is the probability that a user will eventually view the video shared by his/her friend. To some extent, both $ShR(t)$ and $ViR(t)$ reflect how interesting the video is. $W(t)$ is the number of sharers' friends by time t who have not yet viewed the video. In other words, $W(t) =$ the number of all sharers' friends - $|V(t)|$. Similar to [10], we assume the $W(t)$ users view the video at a constant rate, which is denoted by λ. $f(S(t))$ is the number of friends of the new sharer exclusive of those friends who viewed the video before the time t. Generally, the average new potential viewers brought by per new sharer will decrease as the increase of the number of sharers in $S(t)$, because most of the new sharer' friends may have already viewed the video from his/her other friends who also shared the video earlier than the new sharer.

Based on the above notations, the propagation process of one video can be described by the following three equations:

$$\begin{cases} \dfrac{d|V(t)|}{dt} = \lambda \cdot W(t) & (6) \\ \dfrac{d|S(t)|}{dt} = ShR(t) \cdot \dfrac{d|V(t)|}{dt} & (7) \\ \dfrac{dW(t)}{dt} = \dfrac{d|S(t)|}{dt} \cdot f(S(t)) \cdot ViR(t) - \dfrac{d|V(t)|}{dt} & (8) \end{cases}$$

where Eq. 6 reflects that the increased viewers during the time dt come from the potential viewers $W(t)$, who are going to view the video at a rate of λ. Eq. 7 reflects that $ShR(t)$ portion of new viewers (d|V(t)|) will become sharers during the time dt. Based on the previous measurement work [5], here we assume that viewers will immediately share the video after the viewing, otherwise will never share the video. Recalling that we define $W(t)$ as the number of all sharers' friends - $|V(t)|$. Accordingly, the variation of $W(t)$ during time dt (dW(t)) can be expressed as the combination of the growth in the number of potential viewers brought by new sharers ($d|S(t)| \cdot f(S(t)) \cdot ViR(t)$) and the reduction caused by the views during dt ($-d|V(t)|$). This relation is given in Eq. 8.

Initially, there is only one sharer (we call it *initiator*), who posted the video from a VSS. Thus, $S(0)$=1, $V(0)$=1, and $W(0)$ is equal to the number of friends of the initiator multiplying $ViR(0)$. There are four parameters that will affect the evolution of $W(t)$: ShR, ViR, $f(S(t))$ and λ. ShR and ViR reflect the characteristics of specific videos to some extent; $f(S(t))$ depends on the friends of the sharers and social topology around them; λ depends on the frequencies users visit the OSN and watch videos. Our prediction framework in the following subsections will introduce how these parameters can be extracted from real trace.

For ease of exposition, Table 2 provides a reference for major notations used in this paper. Generally, we use upper superscript k (e.g., k in V^k) to denote a video k, and lower subscript i (e.g., i in V_i) to denote a user i. Note that for concise presentation, sometimes we may omit the video superscripts under the premise of no concept confusion (e.g., use $V(t)$ to denote $V^k(t)$ of video k).

Table 2: Summary of major notations

Notation	Description
F_i	set of the friends of user i;
$V_{i\to j}$	set of videos shared by user i and viewed by user j;
V_i	set of videos viewed by user i;
S_i	set of videos shared by user i;
S_{F_i}	set of videos shared by user i's friends;
ShR_i	the average probability that user i will share the videos that s/he viewed;
$ViR_{i\to j}$	the average probability that user j will view the videos shared by its friend user i;
BrF_i	the average number of friends will view a video shared by user i; $BrF_i = \sum_{j\in F_i} ViR_{i\to j}$;
$V^k(t)$	set of viewers of video k until time t;
$S^k(t)$	set of sharers of video k until time t;
v^k_Δ	number of views of video k during period of Δ
$W^k(t)$	number of waiting viewers of video k at time t
α^k	a factor that reflects the normalized ShR of video k;
β^k	a factor that reflects the normalized ViR of video k;
ShR^k	the average probability video k will be shared after being watched;
ViR^k	the average probability video k will be viewed by a friend of a sharer;
ShR^k_i	probability user i will share video k that s/he viewed;
$ViR^k_{i\to j}$	the probability that user j will view the video k shared by its friend user i;
t^k_i	sharing time of video k by sharer i;
λ	the rate of users counted in $W(t)$ who will view video in current time instance;
$\Phi(t)$	the CDF of time (t) between a share and the viewing from the sharers' friends;
$f(S(t))$	the number of potential viewers brought by a new sharer given $S(t)$;

5.2 Framework of SoVP

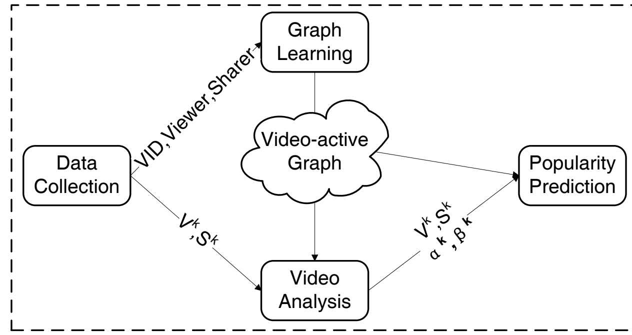

Figure 7: Framework of SoVP

The propagation-based prediction architecture, as shown in Fig. 7, consists of data collection module, graph learning module, video analysis module, and popularity prediction module. First, the data collection module collects logs that record user viewing actions. The basic log format is (Video ID, Viewer ID, Sharer ID, Time), the meaning of which is described in Section 3. Then the logs are taken as the inputs by the graph learning module and the video analysis module. For the graph learning module, historic user viewing records are used as the input. The graph learning module generates a graph called video-active graph, which records the viewing-sharing relationships between users as well as the statistics of user sharing and viewing actions. The video analysis module takes two kind of inputs: video information (sharers S^k and viewers V^k) that is got directly from the data collection module, and the video-active graph that is generated by the graph learning module. The video analysis module analyzes video attractiveness (α^k,β^k) in the context of the video-active graph. Finally, the popularity prediction module uses both the video-active graph and the video attractiveness to make predictions.

5.3 Video-active Graph Learning Module

The topology of an OSN is an important influencing factor to the propagation of videos shared in it. Instead of simply using the original unweighed friend-friend graph, we build a weighted graph called video-active graph. There is a directed edge from user i to user j if the user j ever viewed a video shared by the user i. We assign weights to vertices and edges according to users' viewing and sharing activity. Users show inhomogeneous activity in sharing and viewing videos. For example, as shown in Fig. 8, the power-law distribution indicates that the numbers of videos viewed by each user in one-month period exhibits large skewness.

Fig. 9 illustrates the properties of vertices and edges in the video-active graph. The properties of a vertex i include a set of viewed videos (V_i), a set of shared videos (S_i), and sharing rate (ShR_i). The properties of an edge (i, j) include $V_{i\to j}$, which is defined as the set of video viewed by user j and shared by user i, and $ViR_{i\to j}$, which is defined as the ratio that user j has viewed the videos shared by user i. Taking records (Video ID, Viewer ID, Sharer ID) as the input in a chronological order, V_i, S_i, $V_{i\to j}$ can be extracted directly. ShR_i and $ViR_{i\to j}$ can thus be calculated by $ShR_i = \frac{|S_i|}{|V_i|}$, and $ViR_{i\to j} = \frac{|V_{i\to j}|}{|S_i|}$, respectively.

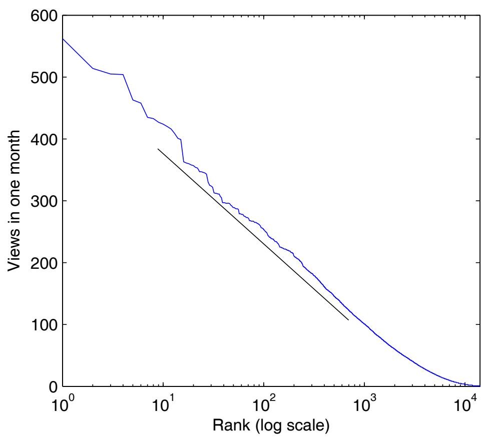

Figure 8: Distribution of user views in one month

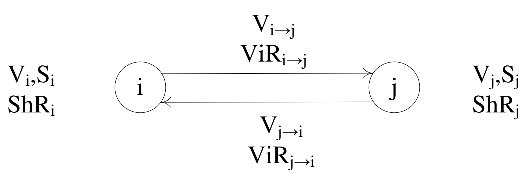

Figure 9: Propeties of the video-active graph

In real OSN systems, the video-active graph grows gradually, continuing to bring new vertices and edges especially at their early stage. Statistics of these newly added edges and vertices cannot be measured directly from real trace at such an early stage. The learning process should adapt to this dynamics. For a new friend link created between two users i and j, time is needed for the $ViR_{i\to j}$ be learned from the interaction between the two users. As such, it is necessary to estimate it from the relationships between i, j and their friends F_i, F_j. We denote the estimated value as $\widehat{ViR}_{i\to j}$, and use Eq. 9 to calculate its value:

$$\widehat{ViR}_{i\to j} = \frac{|V_j|}{|S_i \cap S_{F_j}|} \tag{9}$$

where V_j is the set of videos that are viewed by the user j; S_i is the set of videos that are shared by the user i; S_{F_j} is the set of videos that are shared by the user j's friends. We take $\widehat{ViR}_{i\to j}$ as the initial value for $ViR_{i\to j}$.

5.4 Video Analysis Module

For a given video k, the video analysis module uses the video statistics (V^k, S^k) provided by the data collection module to analyze its attractiveness in the context of the video-active graph. Both ShR and ViR are influenced by the video's attractiveness as well as the characteristics of involved users, so that they are not suitable be used to exactly reflect a video's attractiveness. For example, one video is shared among the users who are very active to share and watch videos, while another video is shared among the users with less activeness. The two videos may happen to have same ShR and ViR based on the simplest definition. Therefore, to gain real values of a video's attractiveness, the video analysis module should remove the effect of the involved users.

For the video k, the video analysis module calculates two factors ($\alpha^k(t)$ and $\beta^k(t)$) to reflect the normalized video attractiveness. The calculation methods are shown in Eq. 10 and 11, respectively.

$$\alpha^k(t) = \frac{|V^k(t)|}{\sum_{i\in S^k(t)}(\Phi(t - t_j^k) \cdot \sum_{j\in F_i} ViR_{i\to j})} \tag{10}$$

where $\Phi(t)$ is the cumulative distribution function (CDF) of time span between sharing a video and the actual view of this shared video by the sharer's friends. We studied the fitting function in the prior work [5]. It is a combined distribution with Weibull (t≤2100, k=0.392, λ=1945) and Generalized Pareto (x≥2100, μ=-2654, σ=6315, ξ=0.669) [5]. t_j^k is the sharing time of video k by sharer j. $|V^k(t)|$ is the actual number of cumulated viewers of video k by time t. $\sum_{i\in S^k(t)}\sum_{j\in F_i}(ViR_{i\to j}\cdot\Phi(t))$ is the estimated average number of cumulated viewers over all videos. The α of an attractive video is usually bigger than 1.

$$\beta^k(t) = \frac{|S^k(t)|}{\sum_{i\in V^k(t)} ShR_i} \tag{11}$$

where $|S^k(t)|$ is the actual number of cumulated sharers of video k by time t. $\sum_{i\in V^k(t)} ShR_i$ is the estimated average number of cumulated sharers over all videos. The β of an attractive video is usually bigger than 1.

When making predictions, we use Eq. 12 and Eq. 13 to decide whether a user will view or share the video k, respectively. The decisions depend on both the video attractiveness and social context.

$$ViR^k_{i\to j} = \alpha^k(t)\cdot ViR_{i\to j} \tag{12}$$

$$ShR^k_i = \beta^k(t)\cdot ShR_i \tag{13}$$

5.5 Popularity Prediction Module

Based on our propagation model, the popularity prediction module takes the information of both video attractiveness and the video-active graph as the input to make predictions.

We rewrite Eq. 6 as Eq. 14, which calculates the number of video views during the time Δ (e.g., one day in this paper). And v_Δ is what we finally need to calculate to be as the predicted views during the time Δ. According to Eq. 14, we need $W(t)$ to calculate v_Δ. We can easily calculate the $W(t)$ at the beginning time of Δ by Eq. 15. Then what we also need to do is to infer $W(t)$ during the time Δ.

$$v_\Delta = |V(T+\Delta)| - |V(T)| = \int_T^{T+\Delta} \lambda\cdot W(t)\, dt \tag{14}$$

$$W(T) = \sum_{i\in S^k(T)}\sum_{j\in F_i} ViR^k_{i\to j} - |V(T)| \tag{15}$$

From Eq. 6, 7, and 8, we get Eq. 16.

$$\frac{dW(t)}{dt} = \lambda\cdot W(t)\cdot(ShR(t)\cdot ViR(t)\cdot f(S(t)) - 1) \tag{16}$$

We define ω as:

$$\omega = \lambda(ShR(t) \cdot f(S(t)) \cdot ViR(t) - 1) \tag{17}$$

Then Eq. 16 can be rewritten as Eq. 18.

$$\frac{dW(t)}{dt} = \omega \cdot W(t) \tag{18}$$

Since in a short period the users' interest in a video will not vary a lot, we assume ω is a constant value from time T to $T + \Delta$, Eq. 18 can be further expressed as Eq. 19.

$$W(t) \approx \delta \cdot e^{\omega \delta t} \tag{19}$$

where δ can be calculated using the initial value of $W(t)$ at time T, as is shown in Eq. 15.

Finally, from Eq. 14 and 19, we get:

$$v_{\Delta} = |V(T + \Delta)| - |V(T)| \approx \frac{\lambda}{\omega}(e^{\omega\delta(T+\Delta)} - e^{\omega\delta T}) \tag{20}$$

where T and $T + \Delta$ are the beginning time and the end time of the day when we need to predict.

6. PERFORMANCE EVALUATION

In this section we compare the performances of conventional views-based prediction models with our propagation-based prediction model, SoVP. We first examine their overall performance on a large set of popular videos. We further examine their performances on the three typical popular videos, which can provide a direct illustration about what kind of evolutions may make the conventional prediction models inefficient.

6.1 Performance Metrics

We evaluate the efficiency of the prediction models using the metric of Relative Absolute Error (RAE). For the video k on the day t, we have:

$$RAE_k(t) = \frac{|\widehat{N}_k(t) - N_k(t)|}{N_k(t)} \tag{21}$$

where $\widehat{N}_k(t)$ is the predicted number of views of video k on the day t, and $N_k(t)$ is the actual number of views. For the average RAE of all testing videos on the day t, we have:

$$RAE(t) = \frac{\sum_k |\widehat{N}_k(t) - N_k(t)|}{\sum_k N_k(t)} \tag{22}$$

For the average RAE of all testing videos on all testing days, we have:

$$RAE = \frac{\sum_t \sum_k |\widehat{N}_k(t) - N_k(t)|}{\sum_t \sum_k N_k(t)} \tag{23}$$

6.2 Prediction Results

As shown in the previous work [17], video popularity distribution exhibits extremely high skewness that top-2% videos account for over 90% views. For the remaining 98% unpopular videos, any of them only received less than 10 views per day on average. Therefore, we take those top-2% popular videos that were initially shared on the same day (March 24^{th}, 2011) as our test set.

First, we need to select proper models for MLR and kNN. We split our data set into a training set that contains the viewing information of 27000 videos, and a test set that contains the viewing information of another 5000 videos. For both MLR and kNN regression, we vary the value of n from 1 to 9; for kNN regression, we also vary the value of k from 1 to 4. We evaluate the performance of each setting on the test data set and the results are shown in Fig. 10 and 11, respectively. Considering the tradeoff of RAE and complexity, we select $n = 5$ for MLR, and $n = 1$ and $k = 3$ for kNN.

Then, we evaluate the overall performance of SoVP as well as the three conventional models with the selected parameters. The average RAE over all test videos for each day is shown in Fig 12. Overall, the SoVP has much better prediction performance than other three models. It is worth noting that ARIMA requires several (e.g., 4 in our experiments) days of early views to learn the model, and so its prediction starts from the fifth day. For MLR, $n = 5$ is used starting from the sixth day, and smaller values are used for earlier days (e.g., $n = 1$ for the second day and $n = 2$ for the third day). ARIMA works well in later days, say after 12 days. It can dynamically select the length of historical information used to predict for each day. For MLR, it works better during the first 10 days and its performance is rather stable. kNN shows dynamic performance. For some days it has the most accurate prediction while for others it performs much worse. The reason is that only the number of views during the last day is used and the popularity distribution could change significantly day by day.

Table 3: RAE of predictions for the type-1 video

	day 2	day 3	day 4	day 5	day 6
kNN	0.823	0.580	0.765	0.720	0.314
MLR	0.886	0.952	0.907	0.820	0.742
SoVP	0.262	0.247	0.186	0.208	0.157

Table 4: RAE of predictions for the type-2 video

	day 2	day 3	day 4	day 5	day 6
kNN	2.729	2.386	1.199	0.212	2.659
MLR	0.843	0.811	0.661	0.538	0.233
SoVP	0.179	0.087	0.108	0.129	0.183

Table 5: RAE of predictions for the type-3 video

	day 26	day 27	day 28	day 29	day 30
kNN	0.926	0.920	0.937	0.808	0.932
MLR	0.951	0.942	0.921	0.832	0.805
ARIMA	0.826	0.684	0.947	0.631	0.219
SoVP	0.400	0.525	0.290	0.327	0.429

We also apply prediction models to the three typical videos that are depicted in Section 4. The original daily views as well as the prediction results are shown in Fig. 13, 14, and 15 respectively. Overall, we can see that the predictions of the three conventional models deviate a lot from the real values, while SoVP works much better than other three models, especially when predicting during the request bursting periods. Since views during the short-term bursts usually count

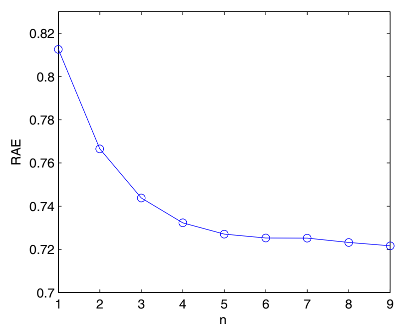

Figure 10: Parameter selection for MLR

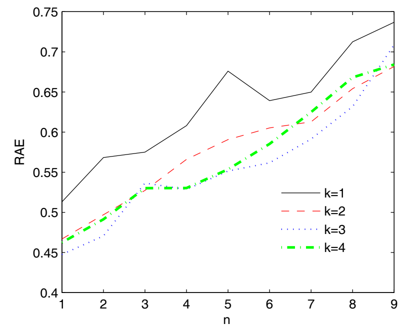

Figure 11: Parameter selection for kNN

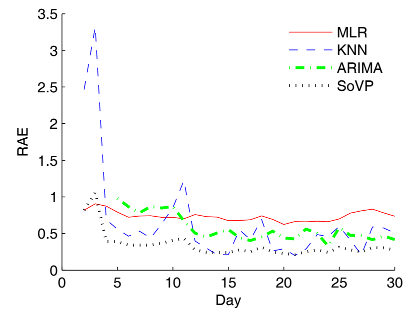

Figure 12: Average performance for testing videos

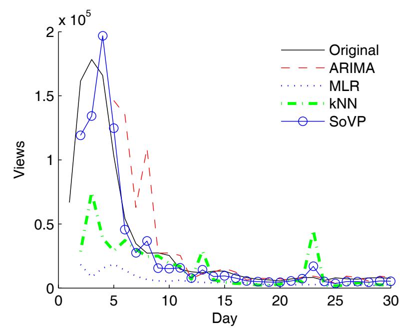

Figure 13: Type-1 video prediction

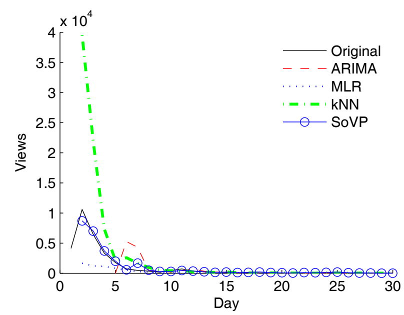

Figure 14: Type-2 video prediction

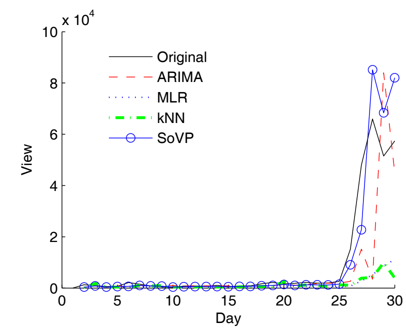

Figure 15: Type-3 video prediction

for most proportion of the video's life-time views, we further give the RAEs of the four models during three videos' bursting days, in Table 3, 4, and 5 respectively. It confirms our observations in the figures. While some further optimizations can be made on those views-based models, they have inherent limits in predicting views with highly dynamic evolution. Solely based on early views, they have difficult to judge a video's sudden increase or decreases in views from its own early evolution pattern, or learning from other early published videos. By contrary, SoVP knows exactly the video's propagation process in the OSN and can extract useful statistics, so that can easily judge whether a video is on increasing stage or decreasing stage, and how fast of this trend.

7. CONCLUSIONS AND FUTURE WORK

This paper presented an initial study on popularity prediction of videos shared in OSNs. We measured and analyzed the characteristics of video propagation and popularity in a large-scale Facebook-like OSN. The results suggested that the video views in early and later times exhibits much less correlation than that in VSSes, which poses significant challenge on conventional views-based prediction models. Our experiments with such conventional prediction models as ARIMA, MLR, and kNN confirmed their ineffectiveness in this new context, especially when predicting the requests bursts that are common for the evolutions of videos shared in OSNs. To overcome the limits, we developed a dynamic model to analyze the video propagation process, and accordingly presented a propagation-based prediction framework, SoVP. SoVP considers both video attractiveness and social context in predicting future video views, whose accuracy has been demonstrated by our trace-driven experiments.

Although SoVP can generally get better prediction than the conventional views-based prediction models, its complexity and scalability are not as good as them. Therefore, a compromised solution between SoVP and the conventional models may be a better choice, and we will consider it in our future work. For example, one possible solution could be simplifying SoVP by only leveraging recent video propagation information. We cloud also incorporate the variables used in SoVP into the conventional models.

8. ACKNOWLEDGMENTS

This research is partially supported by a Canadian NSERC Discovery Grant, a Swedish STINT Initial Grant, a Chinese NSFC Major Program of International Cooperation Grant (61120106008), a Chinese NSFC Project (61170292), and a Start-up Grant from the University of Mississippi.

9. REFERENCES

[1] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proc. of WWW*, 2012.

[2] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. B. Moon. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proc. of IMC*, 2007.

[3] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proc. of WWW*, 2009.

[4] V. Chaoji, S. Ranu, R. Rastogi, and R. Bhatt. Recommendations to boost content spread in social networks. In *Proc. of WWW*, 2012.

[5] X. Cheng, H. Li, and J. Liu. Video sharing propagation in social networks: Measurement, modeling, and analysis. In *Proc. of INFOCOM mini-conference*, 2013.

[6] comScore. http://www.comscore.com/insights/press_releases/2012/8/comscore_releases_july_2012_us_online_video_rankings.

[7] R. Crane and D. Sornette. Viral, quality, and junk videos on youtube: Separating content from noise in an information-rich environment. In *AAAI Spring Symposium*, 2008.

[8] W. Galuba, D. Chakraborty, K. Aberer, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs. In *Proc. of WOSN*, 2010.

[9] G. Gĺźrsun, M. Crovella, and I. Matta. Describing and forecasting video access patterns. In *Proc. of INFOCOM*, 2011.

[10] T. Hogg and K. Lerman. Stochastic models of user-contributory web sites. In *Proc. of ICWSM*, 2009.

[11] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proc. of WWW*, 2011.

[12] F. Kooti, W. A. Mason, K. P. Gummadi, and M. Cha. Predicting emerging social conventions in online social networks. In *Proc. of CIKM*, 2012.

[13] J. G. Lee, S. Moon, and K. Salamatian. An approach to model and predict the popularity of online contents with explanatory factors. In *Proc. of Web Intelligence*, 2010.

[14] K. Lerman and R. Ghosh. Information contagion: an empirical study of the spread of news on digg and twitter social networks. In *Proc. of ICWSM*, 2010.

[15] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *Proc. of WWW*, 2010.

[16] H. Li, J. Liu, K. Xu, and S. Wen. Understanding video propagation in online social networks. In *Proc. of IWQoS*, 2012.

[17] H. Li, H. Wang, J. Liu, and K. Xu. Video sharing in online social network: Measurement and analysis. In *Proc. of NOSSDAV*, 2012.

[18] H. H. Liu, Y. Wang, Y. R. Yang, H. Wang, and C. Tian. Optimizing cost and performance for content multihoming. In *Proc. of SIGCOMM*, 2012.

[19] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in network. In *Proc. of KDD*, 2012.

[20] A. Navot, L. Shpigelman, N. Tishby, and E. Vaadia. Nearest neighbor based feature selection for regression and its application to neural activity. In *Proc. of NIPS*, 2006.

[21] D. Niu, Z. Liu, and B. Li. Demand forecast and performance prediction in peer-assisted on-demand streaming systems. In *Proc. of INFOCOM*, 2011.

[22] H. Pinto, J. Almeida, and M. Goncalves. Using early view patterns to predict the popularity of youtube videos. In *Proc. of WSDM*, 2013.

[23] J. L. Rodgers and W. A. Nicewander. *Thirteen ways to look at the correlation coefficient*. The American Statistician, 1988.

[24] T. Rodrigues, F. Benvenuto, M. Cha, K. P. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. In *Proc. of IMC*, 2011.

[25] M. Rowe. Forecasting audience increase on youtube. In *Proc. of the International Workshop on User Profile Data on the Social Semantic Web*, 2011.

[26] D. A. Shamma, J. Yew, L. Kennedy, and E. F. Churchill. Viral action: Predicting video view counts using synchronous sharing behaviors. In *Proc. of ICWSM*, 2011.

[27] S. Siersdorfer, S. Chelaru, and J. S. Pedro. How useful are your comments? analyzing and predicting youtube comments and comment ratings. In *Proc. of WWW*, 2010.

[28] G. V. Steeg, R. Ghosh, and K. Lerman. What stops social epidemics? In *Proc. of ICWSM*, 2011.

[29] E. Sun, I. Rosenn, C. Marlow, and T. Lento. Gesundheit! modeling contagion through facebook news feed. In *Proc. of ICWSM*, 2009.

[30] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 2010.

[31] Z. Wang, L. Sun, X. Chen, W. Zhu, J. Liu, M. Chen, and S. Yang. Propagation-based social-aware replication for social video contents. In *Proc. of ACM Multimedia*, 2012.

[32] Z. Wang, L. Sun, C. Wu, and S. Yang. Guiding internet-scale video service deployment using microblog-based prediction. In *Proc. of ICWSM*, 2012.

[33] X. Wei, J. Yang, and L. A. Adamic. Diffusion dynamics of games on online social networks. In *Proc. of WOSN*, 2010.

[34] T. Wu, M. Timmers, D. D. Vleeschauwer, and W. V. Leekwijck. On the use of reservoir computing in popularity prediction. In *Proc. of ICCGI*, 2010.

[35] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li. Citation count prediction: Learning to estimate future citations for literature. In *Proc. of CIKM*, 2011.

[36] YouTube. http://www.youtube.com/t/press_statistics.

[37] L. Zhang, T. Peng, Y. Zhang, and X. Wang. Content or context: Which carries more weight in predicting popularity of tweets in china. In *Proc. of WAPOR*, 2012.

[38] R. Zhou, S. Khemmarat, and L. Gao. The impact of youtube recommendation system on video views. In *Proc. of IMC*, 2010.