Modeling and Predicting the Popularity of Online Contents with Cox Proportional Hazard Regression Model

Jong Gun Lee, Sue Moon, Kavé Salamatian

Modeling and Predicting the Popularity of Online Contents with Cox Proportional Hazard Regression Model

Jong Gun Lee[1,*]

LIP6-CNRS, Université Pierre et Marie Curie, 4 Place Jussieu, 75005, Paris, France

Sue Moon

Computer Science Department, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, 305-701, Republic of Korea

Kavé Salamatian

LISTIC - Polytech Annecy-Chambéry, BP 80439, 74944 Annecy le Vieux Cedex, France

Abstract

We propose a general framework which can be used for modeling and predicting the popularity of online contents. The aim of our modeling is not inferring the precise popularity value of a content, but inferring the likelihood where the content will be popular. Our approach is rooted in survival analysis which deals with the survival time until an event of a failure or death. Survival analysis assumes that predicting the precise lifetime of an instance is very hard but predicting the likelihood of the lifetime of an instance is possible based on its hazard distribution. Additionally we position ourselves in the standpoint of an external observer who has to model the popularity of contents only with publicly available information. Thus, the goal of our proposed methodology is to model a certain popularity metric, such as the lifetime of a content and the number of comments which a content receives, with a set of explanatory factors, which are observable by the external observer.

Among various parametric and non-parametric approaches for the survival analysis, we use the Cox proportional hazard regression model, which divides the distribution function of a certain popularity metric into two components: one which is explained by a set of explanatory factors, called risk factors, and another, a baseline survival distribution function, which integrates all the factors not taken into account. In order to validate our proposed methodology, we use two datasets crawled from two different discussion forums, `forum.dpreview.com` and `forums.myspace.com`, which are one of the largest discussion forum dealing various issues on digital cameras and a discussion forum provided by a representative social networks. We model two difference popularity metrics, the lifetime of threads and the number of comments, and we show that the models can predict the lifetime of threads from `Dpreview` (`Myspace`) by observing a thread during the first 5∼6 days (24 hours, respectively) and the number of comments of `Dpreview` threads by observing a thread during first 2∼3 days.

Keywords: Popularity of online contents, survival analysis, Cox proportional hazard regression model

1. Introduction

The emergence of Web 2.0 and Online Social Networking (OSN) services, such as `Digg`[2], `YouTube`[3], `Facebook`[4], and `Twitter`[5], has changed how users generate and consume online contents. The YouTube site report of 24 hours worth of video upload every minute [6] witnesses that the amount of user-generated contents is growing fast. Sharing and commenting on other users' contents via online social networking services constitute a significant part of today's Internet users' web experience. In this context understanding how do users find contents that are interesting? and how do certain contents rise in popularity? becomes of utmost importance. Such a mechanism will be extremely expedient in this age of information deluge both for providers and users that can privilege those mostly likely to get popular content.

The popularity of an online content is not a well-defined term. It is highly subjective and can be related to a mixture of endogenous and exogenous factors. The choice of factors varies among web services, persons and contents. Nevertheless, the popularity of online contents is by nature difficult to predict. For example, the flurry of contents and reactions happening very early in occasion of the death of Michael Jackson was probably far more than what could have been predicted by any model. Some contents become increasingly popular over

*Corresponding author
Email addresses: `jonggun.lee@lip6.fr` (Jong Gun Lee), `sbmoon@kaist.edu` (Sue Moon), `kave.salamatian@univ-savioe.fr` (Kavé Salamatian)
[1]Telephone and fax: +33.(0)1.44.27.88.77 and +33.(0)1.44.27.53.53
[2]`http://www.digg.com`
[3]`http://www.youtube.com`
[4]`http://www.facebook.com`
[5]`http://www.twitter.com`
[6]`http://www.youtube.com/t/fact_sheet` (accessed on Nov 1, 2010)

time following a cascading effect (1) and it is hard to predict which contents will eventually instigate such a cascading effect. It is also noteworthy that accessibility and observability of the predictor factors may vary; an OSN user has not the same visibility into OSN metrics than an OSN operator. The predictability of online contents will depend on which popularity metrics we have access to. It is therefore important to consider what factors do we take into consideration and which explicit data and related measures could be used to represent the popularity. Last but not least the popularity data are censored by nature; when one evaluates the popularity of an online content at time T, he can never be sure that this content will not continue to attract popularity in the future, so that the prediction done on a content popularity might be attained later. All these difficulties compound the challenge of predicting popularity.

In (2), Szabo and Huberman presented a linear regression to predict the long time popularity of an online content from early measurement of user access pattern. More precisely, the authors of (2) observed a linear correlation between the logarithmically transformed long time popularity of a content with its the logarithm of its early measured popularity, *i.e.* they observe that the order of magnitude of the long time popularity can be predicted using the order of magnitude of early popularity. However, in order to obtain this linear correlation they had to remove from their analysis 11% of contents assumed as outliers. A more annoying fact is that because of the logarithmic transform applied to the popularity, the prediction errors behave as a multiplicative coefficient to the long-term popularity. This results in large prediction errors. The weakness of linear regression and of log-linear regression is also confirmed in (3), where a reservoir computation based on the prediction of the logarithm of the long-term popularity is proposed. However the reservoir computing predictor lacks of explanatory ability, so it is not easy to compare the effectiveness of different predictive factors on the long-term popularity.

In this paper, we take the standpoint of an individual user who has to infer the popularity of a content from publicly observable data, such as the lifetime of threads and the number of published comments. Based on OSN privacy rules, different users may have different views of popularities and their choices of measures would thus be different. Our goal in this paper is to propose a generic approach allowing users to weight the impact of different contributing factors and to choose the few most relevant. Our approach in this paper differs from (2) and (3); rather than targeting a precise prediction of the popularity that as explained will be anyway very difficult, we have an explanatory factor analysis, wish to determine, and weight explanatory factors that could explain the popularity of online contents. Our approach is related to survival analysis in biostatistics. A patient with a cancerous metastasis might stay alive much longer than predicted by his (her) doctors, when a healthy young person might die in a car accident. Nevertheless, predicting the likelihood that one will survive longer than a threshold or another individual is possible. In particular one can evaluate the effect of risk factors (like smoking, blood pressure, or more simply age) and compare their impact. We do not aim at inferring the precise popularity of a content but rather determine the likelihood or the probability that a content with given characteristics will attract popularity above a given threshold. We want also to be able to compare the explanatory power of different predictors in order to choose the most useful variables.

To address the above described goals, we will use in this paper a Cox proportional hazard regression model (4). This approach is frequently applied in biostatistics to model human survival and in reliability theory. This model works on the empirically observed Cumulative Distribution function (CDF) of the content popularity rather than working on the individual popularity values. It divides the CDF of the observable content popularity measure into two components: (a) one that can be explained by the given set of explanatory factors, called risk factors, and (b) a baseline distribution function that integrates all the effect and factors not taken into account by explanatory factors. We motivate later this modeling choice

We validate the use of this model over two datasets crawled from two online thread-based discussion forums: `dpreview.com` and `myspace.com`. Our datasets contain information about 267,000 threads and 2.5 million comments. The use of thread-based OSN is a particularity of this paper that differentiate it further from other works in the literature like (2; 3). Indeed, defining the popularity of a thread is more difficult as generally discussion forums do not provide any information about the content access statistics. We therefore assume that the popularity of a thread is captured by the number of comments in it and by its lifetime.

The contributions of our paper are:

1. We show that the survival analysis is applicable to model predict the popularity of an online content. For this purpose we have implemented the Cox proportional hazard regression model to predict the likelihood of a popularity metric using a set of publicly observable explanatory variables.
2. The survival analysis extracts from the popularity observation what can be related to the given explanatory factors from what cannot be related to them and should be assumed as coming from other sources not considered (or not observable) in the given explanatory factors (the baseline hazard). In particular we show that the baseline hazard can be modeled very well using a Weibull distribution, providing therefore a complete parametric model for describing the popularity of the online contents. Moreover the analysis enables us to weight the different possible explanatory factors and to choose a subset of them for modeling and predicting the popularity.
3. We validate our approach by modeling two kinds of popularity metrics, the lifetime of threads and the number of comments, with two different online discussion forums and show that our proposed approach is able to predict the likelihood of the fate of an online content after only a short period of observation.

The remainder of this paper is structured as follows. In Section 2 we explain survival analysis and the Cox proportional hazard regression model and in Section 3 and 4 we describe our prediction methodology. Section 5 gives our experiment

results of predicting the lifetime of threads and the number of comments of threads. We present related work in Section 6 and finally we conclude this paper in Section 7.

2. Background

In this section we briefly explain the survival analysis and its three key functions in Section 2.1 and describe the Cox proportional hazard regression model and its interpretation in Section 2.2. Then we present how to interpret a set of distributions from two components of the Cox model in a figure in Section 2.3 and we explain why we use survival analysis and Cox proportional hazard regression model to model and predict the popularity of online contents in Section 2.4.

2.1. Survival Analysis

Survival analysis is a branch of statistics that deals with survival time until an event of failure or death. It is widely used in various areas, such as biology, engineering, economics, and sociology. However the approach is generic and can be applied to any random variable. We define three main functions, failure, survival, and hazard functions relevant to survival analysis. As we are applying survival analysis to another domain where the concept of death or failure is irrelevant, we need to define our terms. In the forthcoming and throughout this paper, T represents a random variable denoting the time until an event happens (for example an online content receive is last attention by its audience). We will name this event a "death" or a "failure" event; the value t the wall clock time.

2.1.1. Failure function F(t) and Survival function S(t)

The failure function $F(t)$ or Cumulative Distribution Function (CDF) of the random variable T, is the probability to fail or die before a certain time t. It is defined as $F(t) = Pr\{T \leq t\}$. We moreover define the Probability Distribution Function (PDF), $f(t)$, as the derivate of $F(t)$, $f(t) = \frac{\partial F(t)}{\partial t}$.

These definitions can be extended to the discrete case, where rather than time to die, we are interest in the number of events before dying. When the random variable K represents the number of events, $F(k)$ represents the probability that the instance fails before observing k events. So the PDF, $f(k)$, will be defined as:

$$\begin{cases} f(0) = F(0) \\ f(k) = F(k) - F(k-1), \quad \forall k \geq 1 \end{cases} \tag{1}$$

The survival function $S(t)$ is the probability of survival up to a certain time t or the Complementary Cumulative Distribution Function (CCDF) of T,

$$S(t) = 1 - F(t) = Pr\{T > t\} \tag{2}$$

The survival function has the following remarkable property:

$$\int_0^\infty S(t)dt = \mathbb{E}\{T\} \tag{3}$$

In other words, the surface under the survival function curve gives the mean lifetime.

2.1.2. Hazard function - h(t)

The hazard function $h(t)$ gives the failure rate at time t conditioned on the instance being still alive at time t, *i.e.*, the expected number of failures happening at or close to time t. The hazard function is given by Equation (5):

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr\{t \leq T < t + \Delta t \mid T \geq t)\}}{\Delta t} \tag{4}$$

$$= \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} \tag{5}$$

One can define the cumulative hazard, denoted as $H(t)$, as the overall number of failures that are expected to happen up to time t. The cumulative hazard $H(t)$ is related to the survival function through the below relation:

$$H(t) = \int_0^t h(u)\,du = -logS(t) \tag{6}$$

This results in this essential relationship between the cumulative hazard function and its survival function:

$$S(t) = e^{-H(t)} \tag{7}$$

Additionally, for the discrete case, $h(t)$ is replaced by $h(k)$ defined as :

$$h(k) = \frac{f(k)}{S(k)} = \left(1 - \frac{S(k-1)}{S(k)}\right) \tag{8}$$

2.1.3. Example: the Weibull distribution

We illustrate these three functions with a Weibull distribution (5) in Figure 1. A random variable T following a Weibull distribution with shape factor γ, and scale factor λ will have

$$f(t) = \frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1} \exp\left(-\left(\frac{t}{\lambda}\right)^{\gamma}\right), \; t > 0, \gamma, \lambda > 0 \tag{9}$$

$$F(t) = 1 - \exp\left(-\left(\frac{t}{\lambda}\right)^{\gamma}\right) \tag{10}$$

$$\mathbb{E}\{T\} = \lambda\Gamma(1 + \frac{1}{\gamma}) \tag{11}$$

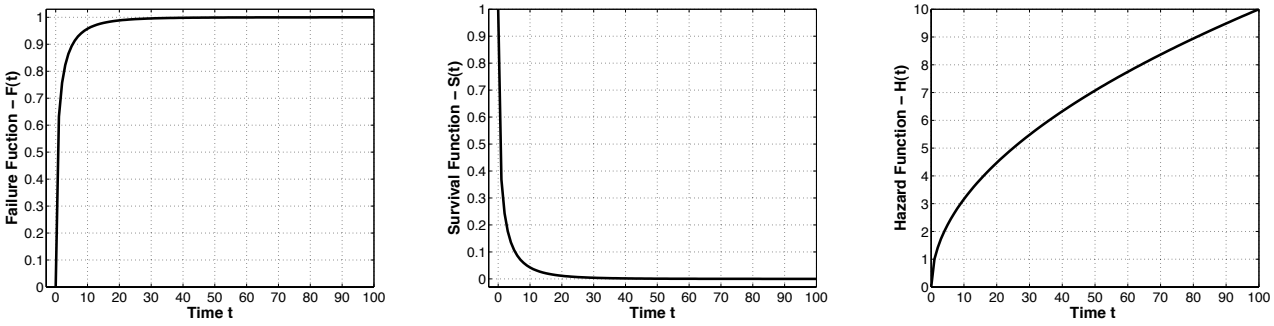
$$\mathbb{V}\text{ar}\{T\} = \lambda^2\left(\Gamma(1 + \frac{2}{\gamma}) - \Gamma(1 + \frac{1}{\gamma})\right) \tag{12}$$

$$h(t) = \frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1} \tag{13}$$

$$H(t) = \left(\frac{t}{\lambda}\right)^{\gamma} \tag{14}$$

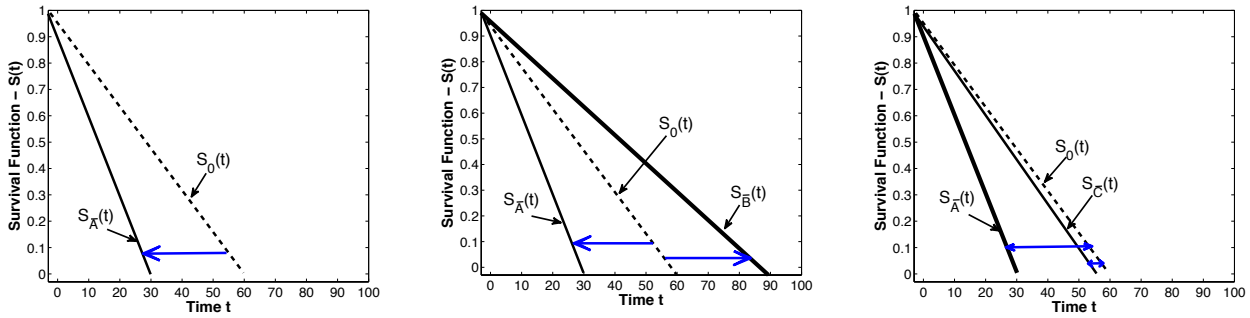
where $\Gamma(.)$ is the gamma function that extends the factorial function to non-integers.

What makes the Weibull distribution remarkable and therefore widely used in survival analysis is the simple polynomial behavior of the cummulative hazard $H(t)$ and the hazard rate $h(t)$. This means that the death rate of a set of Weibull distributed random variables can be only related to time and will depends on only two factors α and λ. In particular, when $\gamma = 1$, $f(t)$ becomes an exponential distribution, the hazard rate become constant, $h(t) = \frac{1}{\lambda}$ and the cumulative hazard grows linearly, $H(t) = \left(\frac{t}{\lambda}\right)$. Under this condition, the constant hazard



(a) Failure function - F(t) (b) Survival function - S(t) (c) Cumulative hazard function - H(t)

Figure 1: Examples of survival, failure, and hazard functions with a Weibull distribution



(a) Base survival function (b) RFs to increase/decrease lifetime (c) More significant risk factor

Figure 2: Examples for understanding risk factors (RFs) and survival function

rate means the remaining lifetime of an alive individual for a duration L does not depends on its age. When $\gamma > 1$, the dying rate is increased with time, *i.e.* the longer the variable lives, the more likely it will die soon. In other words, for $\gamma > 1$, the variable dies from oldness. When $\gamma < 1$, the dying rate is decreased with time, *i.e.* the longer the variable lives, the more likely it will continue to live.

2.2. Cox Proportional Hazard Regression Model

Cox proportional hazard regression (4) is a semi-parametric approach widely used for the survival analysis in practice. In the forthcoming, we describe the Cox regression model given that the failure time is a continuous random variable but the model can be extended in a straightforward way to the case where the failure time is a discrete variable K.

The Cox proportional hazard model assumes that the hazard rate function can be represented as a parametric linear combination of a set of risk factors

$$h(t) = h_0(t) \times exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k). \qquad (15)$$

The hazard function $h(t)$ is composed of two components:

1. A parametric part that depends linearly on the risk factors. The risk factors (RFs) $X = \{x_1, ..., x_k\}$ are the set of factors that influence the survival duration. As the risk factors are introduced through an exponential function, their effects become proportional, *i.e.*, adding to the risk factor has a multiplicative effect on the hazard function. Therefore, the coefficient β_i represents the relative importance of risk factors.
2. The non-parametric part defined as baseline hazard $h_0(t)$ gives the natural risk. This function gives the hazard when any risk factor is not presented. No assumption is made about the form of $h_0(t)$ and its relation with time. Using Eq. 6, one can define a baseline survival function as $S_0(t) = \exp(-\int_0^t h_0(t)dt)$. The baseline survival $S_0(t)$ defines the CCDF describing the death time of a random variable that has not any risk factor in the set X.

The quantity $\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$, called the prognostic value, is not dependent on time. It is used for evaluating the increase (or decrease) in risk with respect to the baseline hazard that an individual under study has to die.

2.3. Interpretation of fitting results

We illustrate the interpretation of the risk factor in Cox proportional hazard regression with three examples in Figure 2. We assume that $S(t)$, represented by the dotted line, is an empirically observed distribution and $S_{\bar{X}}(t)$ is the residual baseline survival function after removing the risk factors $X = \{A, B, C\}$ by fitting the Cox proportional hazard model, *i.e.* $S(t)$ is the survival distribution when the risk factors in X are present and $S_{\bar{X}}(t)$ is the baseline hazard without presence of the risk factors in X. To simplify the discussion, we have presented the survival

functions as straight lines. By definition $S(0) = S_{\bar{X}}(0) = 1$ and they diverge after. The wider the distance between $S(t)$ and $S_{\bar{X}}(t)$ is, the more effective the risk factors in X are.

1. Figure 2(a) shows that the overall lifetime is increased by the effects of the risk factor A as the expected lifetime in presence of A, $S(t)$, is larger than in absence of A, $S_{\bar{A}}(t)$
2. Figure 2(b) shows that while lifetime is increased by risk factor A, it is decreased by B .
3. Figure 2(c) shows the empirically observed lifetime distribution $S(t)$ and two baseline survival functions, $S_{\bar{A}}(t)$ and $S_{\bar{C}}(t)$. We say that the risk factor A is more a significant factor that C because the lifetime in its absence is shorter than the lifetime in abscence of C

Another interesting interpretation is related to Equation 3 and gives a more refined analysis. The surface under the curve of $S(t)$ is the overall mean lifetime, while the surface under the curve of $S_{\bar{A}}(t)$ is the mean lifetime after removing the impact of the risk factor A. We call it the mean baseline lifetime. By observing the difference between these two mean lifetime, one can evaluate the impact of the risk factor on the lifetime; the larger this difference is, the more expressive this risk factor is. This property will be used to propose a heuristic for choosing a subset of explanatory variable for predicting the popularity of online contents.

2.4. Discussion

In this subsection, we will compare Cox proportional hazard regression with other regression techniques and describe why we believe it is well suited to predicting online contents popularity. A review of the existing literature in popularity prediction shows that because of high randomness in empirical data predicting precisely the popularity of online contents has appeared as a difficult problem. One can explain it by the intrinsic unpredictability and sometimes unrationality of popularity. To get around this problem, (2) has applied a logarithmic transformation on popularity, *i.e.* they predicts the logarithm of popularity at the target time with the logarithm of the early popularity. Even after this transform the linear alignment is still not satisfactory and they have to remove 11% of their data (using a two class clustering) as outlier to be able to do their prediction. The authors of (2) do not provide any method for deciding *a priori* if a content will belong to the outlier class or not. Nonetheless, the result model leads to large error on predicting the individual popularity of content as the regression error are magnified by the exponential applied to inverse the effects of the initial logarithm transform, and act as multiplicative coefficients. We show in Figure 2.4 the cloud of points representing the thread lifetime versus the number of comments in the first 6 days in `D-myspace`. As shown, no linear alignment can be seen showing that the simple linear regression is not enough directly to predict the thread lifetime and we have to use another approach.

Whenever, the goals of (2) are similar to ours, the Cox regression is better suited to the goal. With the Cox Regression, we do not try to predict the popularity directly but rather the likelihood

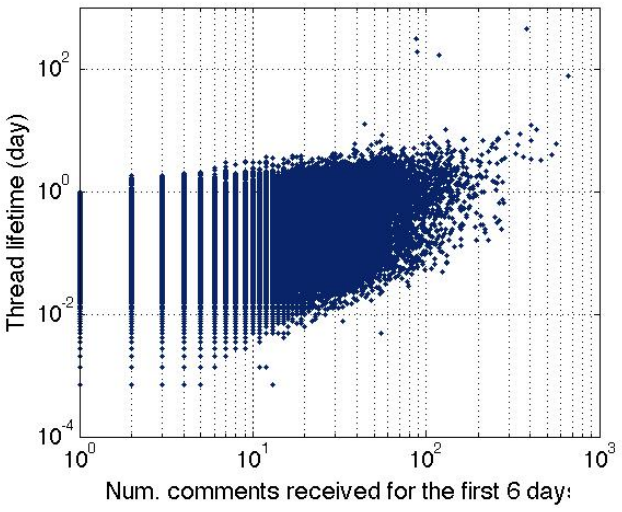


Figure 3: Thread lifetime versus the number of comments in the first 6 days in `D-myspace`

that the popularity cross a threshold, or in other term the probability that a content will survive longer (or attract larger number views or comment) in popularity competition. With this assumption we do not have anymore to remove outliers from our dataset. In fact the outlier effects are described by the baseline hazard that will capture the effect of all factors not taken into account in the risk factors. Whenever the baseline hazard can be modeled by a Weilbull distribution, one can state that baseline hazard is capturing the curse of time (or momentum) effect leading to a very elegant model where the risk factors capture the rational and explanatory factors of popularity and the Weibull baseline hazard that in fact accounts for modeling errors, only depends on time.

In (3), the authors used a reservoir computing neural network approach that achieve reasonable prediction results but their approach consists of predicting the popularity of an online content using a sequence of popularity observed for previous days. This goal is similar to the goal of authors of (6) who apply an ARMA model to predict the popularity of a YouTube content based on the time sequence of past days popularity. Whenever these approaches are interesting they diverge from our goal that is to do a very early prediction of the fate of a content eventually based on a very limited number of observations. Moreover, the authors of (3) state in their conclusions that "*The Reservoir Computing shows excellent prediction performance (...). However, as it is a black box, its prediction capability is difficult to analytically interpret; moreover, because it is usually required to predict the popularity in an early age, the randomization effect still exists in the prediction.*" A strong point of the Cox regression is the fact that the resulting model is very easy to analytically interpret in particular the value e^{β_i}, representing the relative importance of the risk factor i, which can be difficult to extract from a neural network approach. There is moreover a rich literature on interpreting the statistical significance of each

coefficient β_i of the Cox regression in a statistical test theory framework.

The authors of (7) compared decision tree, neural networks, and Support Vector Machines (SVMs) to determine if a content in the Digg cooperative recommendation and filtering OSN will attract a Digg score higher than a target. The goals of this prediction study are similar to our goals. While the prediction results show a good ability to predict the popularity, the approach is missing an interpretative framework that we have with the Cox proportional hazard regression, in particular the ability to compare the effectiveness of different predictive factor.

The above arguments and observation lead us to use the Cox Proportional hazard regression model. However, this model has also some weaknesses which are related to the argument on proportionality; risk factors have a multiplicative effect that is constant throughout the life of the individuals under study. To leverage this condition, authors in the biostatistics literature have suggested to use stratification or time products covariates (8). However we do not need to do this in our case as the fitting results were quite good as the will be described in the forthcoming.

3. Selecting significant risk factors

Similarly to (2), we use as explanatory or risk factors, early values of the online content attributes that are visible to an external user, *e.g.* the number of comments or the number of hits after a short time of the creation of online content. Indeed, this choice is governed by the availability of only these metrics. An OSN operator could have access to richer variable and can use them.

However, the risk factors might be strongly correlated introducing collinearity that is harmful to the quality of inference. We have therefore to select and significant risk factors, instead of using them all. The problem of variable selection in models is a very classical problem. Several generic approaches have been coined to deal with variable choosing. Most of these techniques use penalized likelihood methods that penalize the likelihood with the number of variables used for constructing the model. For example, the LASSO technique uses penalized likelihood method with the L1-penalty (9). The authors of (10) present a non-concave penalized likelihood approach adapted to the Cox proportional hazards model and compare it with the LASSO technique. They show that with a proper choice of the regularization parameter and the penalty function, their proposed estimators are able to choose correctly the best subset of risk factors for regression.

In our case we have only a small number of potential risk factor candidates, so in place of hammering the problem using a penalized likelihood approach, we will describe a visual heuristic based on the description given in section 2.3. This heuristic will also give more insight into the impact of each chosen risk factors.

Our heuristic consists of choosing among the potential risk factors the subset which maximize the performance of the predictive model. This involves a) making all possible combinations of the potential risk factors, b) applying the Cox regression model to each combination, and c) following what described in section 2.3, we calculate the mean baseline lifetime for each combination and evaluate its effect on the mean lifetime. We will choose the subset of potential risk factors that minimize the mean baseline lifetime. We have also to ensure that the potential risk factors are not too correlated to avoid multicollinearity. We ensure this by checking the correlation between every two candidate factors in order to avoid the simultaneous use of highly correlated factors.

4. Cox proportional model fitting

Formally, we model the hazard function $h(t)$ of a popularity metric, through the Cox proportional hazard regression, which we described in Section 2. The popularity metric can be any measure of the popularity of online contents, such as the lifetime of threads and the number of received comments per thread and the risk factors can be any information related to the online contents and observed by a user who wishes to do the prediction of popularity.

After choosing the set of risk factors to be used, one can apply the Cox proportional regression to it and fit the long-term empirical distribution of long-term popularities obtained after the latest activity, such as statistics of comment and view of a content. This fitting can be done using one of the several libraries that implement it.

However, our goal is to apply the Cox proportional regression in order to predict the popularity of an online content as early as possible. We therefore fit different regressions generated using different values of initial observation period. This enables us to find an observation window where the information from risk factors are enough to obtain a good prediction of the likelihood of popular contents. We use the same heuristic as described in section 2.3; the further the baseline hazard goes from the empirical distribution the better becomes the predictive power of the variables relative to this observation period. As explained in section 2, a good fit of the baseline survival function with a Weibull distribution would be of high interest as it will provide an elegant parametric model for the empirical distribution. Thus as the last step we will fit the baseline survival function resulting from the Cox proportional hazard regression, to a Weibull distribution and evaluate the scale and shape parameters, γ and λ. Finally the total hazard function of an empirical observation is inferred as:

$$h(t) = \frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1} \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) \qquad (16)$$

and using Equation (6) one can derive an approximation of the empirical survival distribution. This is noteworthy that the only source of error in this approximation is relative to the fitting of $h_0(t)$ with the Weibull distribution.

The prediction therefore proceed by calculating for a given online content its risk factors, deriving its hazard function that gives the likelihood (through Equation (6)) that the online content will survive longer than a time t.

Dataset	Service address		Forum name	Start - End (Duration)
D-dpreview	http://forum.dpreview.com		Canon 40D-10D	2003/01 ~ 2007/12 (5 years)
D-myspace	http://forums.myspace.com		Music - General	2004/01 ~ 2008/04 (4 years and 4 months)
Dataset	# threads	# comments	# unique users for threads	# unique users for comments
D-dpreview	140,524	1,496,808	27,989	41,269
D-myspace	127,607	1,038,989	-	-

Table 1: Description of our two datasets crawled from two online discussion forums

5. Experimental validation

In this section, we validate our proposed modeling methodology with two datasets crawled from two online discussion forums.

5.1. Datasets description

We collected two datasets, D-dpreview and D-myspace, extracted from online discussion forum services of forums.dpreview.com and forum.myspace.com. The Table 1 shows a brief description and gives some statistics about the two datasets.

The web service Dpreview provides its users discussion forums about the specifications of all kinds of digital cameras. The D-dpreview dataset was made by crawling the information of all threads and all comments related to the Canon EOS 40D-10D. This topic has the largest number of threads when we started our data collection. We collected for each post, either a thread or a comment, it position in the thread hierarchy, its anonymized user identifier and its posting timestamp. As the table shows, we have gathered the information of all posts posted during the five year between 2003 and 2007.

The OSN Myspace is a representative service where users create and share various contents, such as texts, images, and videos, with their social contacts. One of its functionalities is discussion forums where the users discuss and debate different issues categorized both by the service provide and the users. We crawled the information of all threads and comments from the Music-General forum, which had the largest number of threads among all forums, when we initiated our collection. We crawled from the forum the timestamp of each post posted between Jan 2004 to Apr 2008 and its position in the thread hierarchy. Differently from the D-dpreview, we could not collect the anonymized user identifier because the information was hidden by the service provider.

We plan to model describe the popularity of a thread using its lifetime and the number of comments in it. However we need to first define the lifetime of a thread. In the rest of this paper, we define the lifetime of a thread as the time span between the posting timestamp of the thread and the post timestamp of its last comment. Indeed as we have not access to the access statistics of thread contents, a thread can still be accessed and used long after the last comment is posted to it. Moreover it is not possible to definitely decide, if a comment is the last one as new comments might be added after our data collection. This means that our data might be censored. Because of this, we assume that a thread is dead if it does not receive any new comment before its expiration time. We choose the thread expiration delay using the distribution of inter-arrival time of two consecutive comments in a thread. We show in Figure 4 the CDF of inter-arrival time in D-dpreview and D-myspace, respectively. The CCDF in

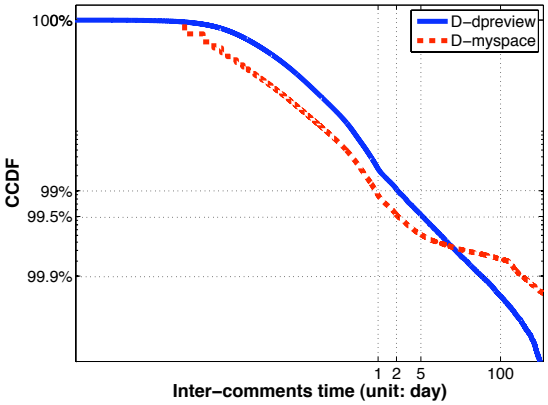


Figure 4: CCDF of Inter-comments time

Figure 4 shows that almost all the inter-comments time is less than 1 day and also shows that about 99.5% of inter-comments time are less than 5 days for D-dpreview and less than 2 days D-myspace. Based on this observation we set the thread expiration delay for D-dpreview as five days (resp. two days for D-myspace).

5.2. Modeling the Lifetime of Threads

In this section, we present the results of modeling the popularity using thread lifetime. As explained we should use explanatory (risk) factors that are visible to an external observer. We assume that a user leaves comments on a thread that he is interested about. Based on this assumption, we have considered two class of potential explanatory factors: global and temporal factors. We choose the following three potential explanatory factors all observable and measured at the observation time:

1.The total number of comments in a thread.

2.The number of comments posted by the author of the thread.

3. The number of unique users beside the author.

The temporal interest is captured in the inter-commenting times. An interesting thread can be expected to receive frequently comments. Thus, we consider also the following information as potential explanatory factors for the temporal user interests.

4. The time until the first comment

5. The median of inter-comments time
6. The mean of inter-comments time
7. The variance of inter-comments time

As we stated before, we have first to filter out useless factors which do not capture effects relative to the thread lifetime. For this purpose, we apply the heuristic described in Section 3. We

Risk factor	1	2	3	4	5	6	7
Mean value	*0.94*	1.38	*1.05*	*0.91*	1.37	*0.93*	*0.85*

Table 2: Mean value of the baseline survival function of the thread lifetime in `D-dpreview` dataset. Each mean value is obtained after removing the effect of an individual risk factor. The overall mean thread lifetime is 1.48

present in table 5 the mean value of the baseline survival function of the thread lifetime derived as the surface under the baseline survival function after removing the effect of an individual risk factor. One can see that variables 2 and 5 are almost useless as these factors can only explain a value 0.1 day out of the overall delay 1.48 day. We therefore remove these two factors from potential explanatory factors and consider only the five remaining explanatory factors:

1. the number of comments,
3. the number of comments by a thread poster,
4. the number of comment contributors,
6. the mean of inter-comment time,
7. the variance of inter-comment time.

In order to go further in the filtering, we checked the correlation coefficient among variables presented in Table 3. The table

RF	1	3	4	6	7
1	1.0000	0.6429	**0.9124**	-0.0004	0.1000
3	0.6429	1.0000	0.4777	0.1000	0.1000
4	**0.9124**	0.4777	1.0000	0.0000	0.1000
6	-0.0004	0.1000	0.0000	1.0000	0.8530
7	0.1000	0.1000	0.1000	0.8530	1.0000

Table 3: Correlation coefficient between two risk factors (RFs). Each number for RF is the same one used in above.

shows that the first explanatory factor is highly correlated with the fourth one. Thus, instead of using both factors, we use only one of them and now we use the first risk factor.

Thereafter we have fitted the Cox regression for all possible combinations of the four remaining explanatory factors, 1, 3, 6, and 7 and derive the surface under the resulting baseline survival functions. The results are shown in Table 5, where it can be seen that among all combinations the best results is obtained with the combination of the four explanatory factors. This combination of explanatory factors can explain 0.92 days (63%) out of the mean thread lifetime of 1,48 days. However a more precise look at table 5 shows that almost the same mean is obtained using explanatory factors 1 (number of comments) and 6 (mean inter-comment time) alone. We therefore decide to use this combination as explanatory factors.

Combination of risk factors	Mean value
(1, 3, 6, 7)	*0.56*
(1, 3, 6)	0.56
(1, 3, 7)	0.58
(3, 6, 7)	0.68
(1, 3)	0.92
(1, 6)	0.57
(1, 7)	0.59
(3, 6)	0.72
(3, 7)	0.69
(6, 7)	0.85

Table 4: Mean value of the fitted Weibull distribution with each combination of risk factors in modeling the thread lifetime (`D-dpreview`)

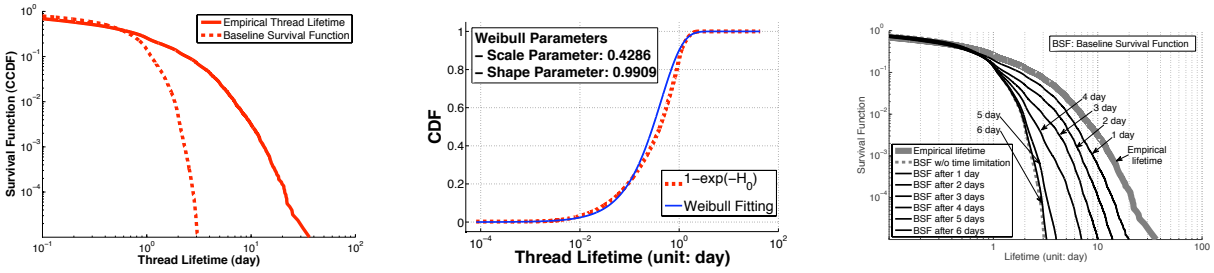
In Figure 5(a) and 5(d), we plot the empirical survival function $S(t)$ and the baseline survival $S_0(t)$ obtained from fitting the Cox regression with four (resp. three risk) factors to `D-dpreview` (resp. `D-myspace`). As described before, we fit the resulting baseline survival $S_0(t)$ with a Weibull distribution and we present the result of the fit in Fig. 5(b) and 5(e). These two figures show that the two baseline survival distributions are well fitted with the Weibull distribution. The obtained scale and shape parameters are $\gamma = 0.9909$ and $\lambda = 0.4286$ for `D-dpreview` (resp. $\gamma = 0.8616$ and $\lambda = 0.101$ for `D-myspace`). The shape factors γ for the two dataset are very close to 1, meaning that the baseline survival can be described as an exponential distribution and therefore a constant death rate of 2.33 (resp. 9.90). This is an important outcome as it means that only the risk factors to explain the variations in the lifetime of different threads as the age is irrelevant.

In Figure 5(c) and Figure 5(f), we investigate if we can model the empirical survival function using the explanatory factors but observed early in the lifetime of the thread. For this purpose we fit a Cox regression model but with explanatory factor observed at different time interval (1 day to 6 days for `D-dpreview` and 3 hours to 24 hours for `D-myspace`) after the creation of the thread. The resulting baseline hazard moves to the left with increasing observation time. Interestingly observing a thread only 6 days over `D-dpreview`, give as much information as observing its explanatory factor over the whole lifetime. Similarly for `D-myspace`, observing a thread 24 hours after its creation give almost as much information about the fate of the thread than observing it explanatory factor over the whole lifetime.

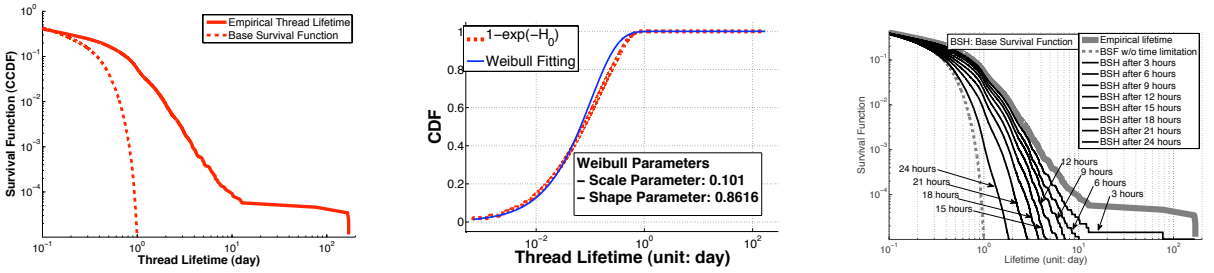
In order to show the predictive power of the explanatory factors, we plot in Figure 6, the thread lifetime as a function of the overall hazard $e^{\sum_i \beta_i x_i}$. It can be seen clearly that a large hazard results in a small lifetime, and inversely all long living threads have small risk. Nonetheless, the relationship between overall risk and thread lifetime is not unique and therefore does not enable a precise prediction, but rather the prediction of the likelihood of the thread lifetime.

5.3. Modeling the Number of Comments in a thread

In this section we extend the analysis done in previous section to the modeling and prediction of the number of comments in a thread. We use the same `D-dpreview` dataset described before, with the same seven potential explanatory factors used in

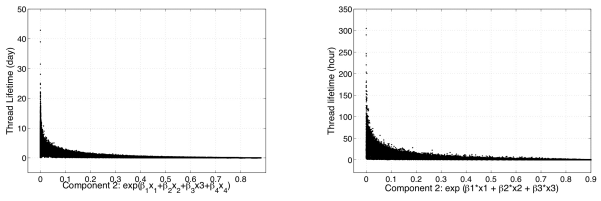


(a) Finding a baseline survival function (b) Fitting it with a Weibull dist. (c) Predicting the lifetime of threads



(d) Finding a baseline survival function (e) Fitting it with a Weibull dist. (f) Predicting the lifetime of threads

Figure 5: Prediction of the lifetime of threads from D-dpreview (upper figures) and D-myspace (lower figures)



(a) D-dpreview (b) D-myspace

Figure 6: risk factor component $(x_1\beta_1 ... x_k\beta_k)$ *vs.* the lifetime of threads

Risk factor	1	2	3	4	5	6	7
Mean value	7.11	10.22	*8.64*	*7.08*	9.14	9.14	9.06

Table 5: Mean value of the baseline survival function of the number of comment in D-dpreview dataset. Each mean value is obtained after removing the effect of an individual risk factor. The overall mean thread lifetime is 10.47. The variable 1 is the number of comments at thread expiration delay

the previous section. However there is a slight issue there. Using the number of comments in a thread for predicting it seems trivial. In fact the practical prediction problem consists of using the early measurement of number of comments in a thread to predict its final value. We therefore will use the number of comments at the thread expiration delay as explanatory (risk) factor. This choice is motivated by the fact that at before thread expiration delay the number of comments is not yet definitive and predicting it is still meaningful.

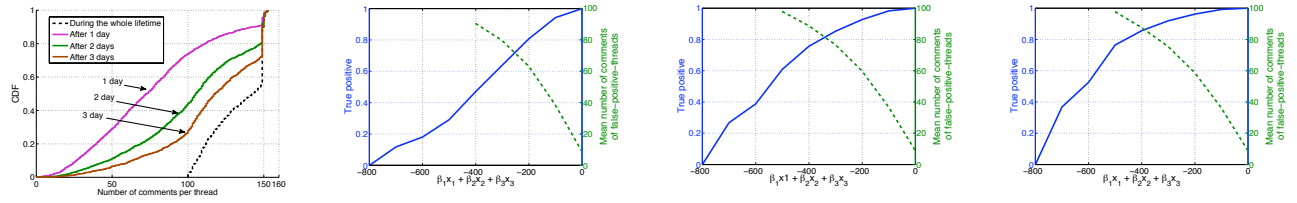
We show in Table the surface under the baseline survival function of the number of comments in D-dpreview dataset. As the overall, mean number of comments is 10.47 clearly, variables 2,5,6 and 7, are not very expressive. Resulting in a choice of explanatory factors as 1,3 and 4. A check on their correlation shows that the correlation between these variables is not large. Applying the heuristic described in section 2.3, we checked all combination of these three explanatory factors. The results are shown in Table 6. This table shows that the variables 3 and 4 are the most effective in explaining the number of content as they can explain 3.84 comments among the mean of 10.47 comments per thread. The parameters of the calibrated Weibull distribution are $\gamma = 1.83$ and $\lambda = 7.8$. We show in Figure 7 the result of the fitting.

Combination of risk factors	Mean value
(1, 3, 4)	*6.61*
(1, 3)	7.21
(1, 4)	7.21
(3, 4)	6.60

Table 6: Mean value of the baseline survival function of the number of comment after removing combination of explanatory factors. The overall mean thread lifetime is 10.47.

Similarly to the thread lifetime modeling we show the effect of the observation time to predict the number of comments per thread. For this, we vary observation time as shown in Figure 8. This figure shows that when we use the information captured during the first 24 hours, the information for risk factors is not enough to predict the number of comments. The baseline survival function using the information observed for more than 2 days, however, is close to the baseline survival function based on the whole observation. Thus, we could closely predict the number of comments of threads after observing the information on risk factors for more than 2 days.

Finally we will illustrate here how we can use the model developed in this paper to predict if a thread will attain more than 100 comments. We found in D-dpreview dataset that there are



(a) CDF (b) after 1 day (c) after 2 days (d) after 3 days

Figure 9: Predicting the threads to have more than 100 comments. (In (b), (c), (d), straight lines are true positive values and dotted lines are mean values of false negative values.)

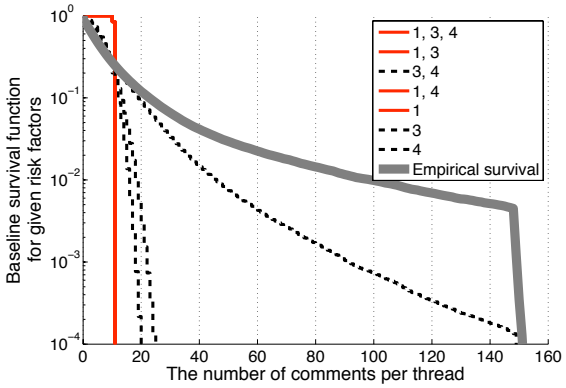


Figure 7: Predicting the number of comments of online discussion forum threads

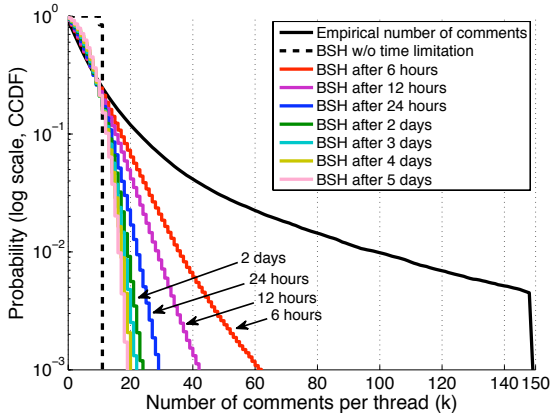


Figure 8: Determining a minimum observation time

1,406 threads which have received more than 100 comments. in Figure 9(a) we plot the number of comments received after one, two, and three days. After one day (resp. two and three days) about 24% (resp. 56% and 73%) of the 1406 threads that have achieved more than 100 comments, will attain 100 comments. Now let's suppose that one observes a thread and calculate a prognostic value of the explicative factor. Now let's predict that the thread will attain the 100 comment if the prognostic value is less than a threshold L. We show in Figures 9(b) to 9(d) the performance of such predictor as a function of the value of the threshold when the prediction is made after 1,2 and 3 days. For example the first day, using a threshold of -200 for prognostic value enable to detect correctly 80% of threads attaining 100 comments. There will be some threads that have been misclassified (false positive threads). The mean number of comments in these thread is equal to 63. Indeed with larger observation time performance improves; the third day, the same threshold enable to detect more than 95% of such thread.

6. Related Work

In this section, we briefly describe the literatures related to our work.

- *Survival analysis*
  Survival analysis (11) has been applied to various areas, such as bio-medical science, sociology, and epidemics (12; 13; 14; 15). Among the methodologies for survival analysis, Cox proportional hazard regression model (4), which is a semi-parametric survival analysis methodology, has been widely used (16; 17; 18). In this paper, we applied the survival analysis and Cox proportional regression approach to model and predict the popularity of online contents.

- *Analysis on Threads and Comments*
  The authors of (19; 20) analyzed the posts and comments of Slashdot. In detail, the authors of (19) explained the behaviors of inter-posts times with statistical models while the authors of (20) analyzed the dynamics of posts and users. (21) characterized the social interactions of Slashdot users by graph theoretic approach. In (22; 23) user comments information wer used to understand user intention. In (24) the detection of influential authors based on user comments was investigated.

- *Modeling Inter-Posting and Popularity prediction*
  The authors of (19) modeled comment time interval with four different statistical models and they predicted intermediate and long-term user activities.In (2) a popularity prediction methodology for online contents was proposed that was based on the correlation of popularity between early and later times. The authors of (25) proposed three prediction models and validated them over Youtube and Digg datasets. In (26), a co-participation network among Digg users was proposed that was used to predict the popularity of online contents using an entropy measure. The

authors of (27) analyzed the relaxation response after the bursty activities caused by endogenous and exogenous factors in Youtube videos. Among all burst activity, the authors that some are followed by a ubiquitous power-law relaxation governing the timing of views. They showed that these bursts could be explained by an epidemic cascade of actions.

Our work is different from the existing literature by our explanatory approach based on Cox proportional hazard model. Rather the predicting the popularity values directly we predict the likelihood that a content will become popular.

7. Conclusion

In this paper, we proposed a methodology for modeling and predicting the popularity of online contents. We applied Cox proportional hazard regression with a set of given explanatory factors to model and predict an objective metric of the popularity of online contents. We validated our approach by modeling and predicting two kinds of popularity metrics: the threads lifetime and the number of comments, with two datasets from two different online discussion forums, `dpreview.com` and `myspace.com`. In our experiments, we, showed that the proposed methodology could predict the lifetime of Dpreview (resp. Myspace) threads by observing a thread during the first 5∼6 days (resp. 24 hours). Moreover, our approach was able to predict the number of comments in Dpreview dataset, by only observing a thread during its first 2∼3 days. Finally, we presented how the approach can be applied to predict the set of threads which would receive more than 100 comments. This last example validates the flexibility of the proposed method.

References

[1] M. Cha, A. Mislove, K. P. Gummadi, A measurement-driven analysis of information propagation in the flickr social network, in: WWW '09: Proceedings of the 18th international conference on World wide web, ACM, New York, NY, USA, 2009, pp. 721–730.

[2] G. Szabo, B. A. Huberman, Predicting the popularity of online content, Communications of the ACM (2010) 80–88.

[3] T. Wu, M. Timmers, D. D. Vleeschauwer, W. V. Leekwijck, On the use of reservoir computing in popularity prediction, in: Proceedings of international conference on evolving internet, IEEE Computer Society, Los Alamitos, CA, USA, 2010.

[4] D. R. Cox, Regression models and life-tables, Journal of the Royal Statistical Society. Series B (Methodological) 34 (2) (1972) 187–220.

[5] W. Weibull, A statistical distribution function of wide applicability, Journal of Applied Mechanics (1951) 293–297.

[6] G. Gursun, M. Crovella, I. Matta, Describing and forecasting video access patterns, in: INFOCOM '11: Proceeding of the 30th IEEE International Conference on Computer Communications, IEEE, 2011.

[7] S. Jamali, H. Rangwala, Digging digg: Comment mining, popularity prediction, and social network analysis, in: Proceedings of international Conference on Web Information Systems and Mining, IEEE Computer Society, Los Alamitos, CA, USA, 2009.

[8] J. P. Klein, Survival Analysis, 2nd Edition, Springer, 2003.

[9] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: A statistical view of boosting, The Annals of Statistics 28 (2000) 337–407.

[10] J. Fan, R. Li, Variable selection for cox's proportional hazards model and frailty model, The Annals of Statistics 30 (2002) 74–99.

[11] R. Schlittgen, Survival analysis: State of the art, Computational Statistics and Data Analysis 20 (5) (1995) 592–593.

[12] A. R. Feinstein, Principles of Medical Statistics, Chapman & Hall/CRC, 2001.

[13] D. G. Kleinbaum, M. Klein, Survival Analysis: A Self-Learning Text (Statistics for Biology and Health), 2nd Edition, Springer, 2005.

[14] A. Diekmann, M. Jungbauer-Gans, H. Krassnig, S. Lorenz, Social status and aggression: a field study analyzed by survival analysis., J Soc Psychol 136 (1996) 761–768.

[15] S. Selvin, Survival Analysis for Epidemiologic and Medical Research (Practical Guides to Biostatistics and Epidemiology), 1st Edition, Cambridge University Press, 2008.

[16] J. Heckman, B. Singer, The identifiability of the proportional hazard model, Review of Economic Studies 51 (2) (1984) 231–41.

[17] P. B. Seetharaman, P. K. Chintagunta, The proportional hazard model for purchase timing: A comparison of alternative specifications, Journal of Business & Economic Statistics 21 (3) (2003) 368–82.

[18] T. M. Therneau, P. M. Grambsch, Modeling survival data: extending the Cox model, Springer, New York, N.Y, 2000.

[19] A. Kaltenbrunner, V. Gomez, V. Lopez, Description and prediction of slashdot activity, in: LA-WEB '07: Proceedings of the 2007 Latin American Web Conference, IEEE Computer Society, Washington, DC, USA, 2007, pp. 57–66.

[20] A. Kaltenbrunner, V. Gómez, A. Moghnieh, R. Meza, J. Blat, V. López, Homogeneous temporal activity patterns in a large online communication space, in: SAW, 2007.

[21] V. Gómez, A. Kaltenbrunner, V. López, Statistical analysis of the social network and discussion threads in slashdot, in: WWW '08: Proceeding of the 17th international conference on World Wide Web, ACM, New York, NY, USA, 2008, pp. 645–654.

[22] M. Hu, A. Sun, E.-P. Lim, Comments-oriented blog summarization by sentence extraction, in: CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, ACM, New York, NY, USA, 2007, pp. 901–904.

[23] B. Li, S. Xu, J. Zhang, Enhancing clustering blog documents by utilizing author/reader comments, in: ACM-SE 45: Proceedings of the 45th annual southeast regional conference, ACM, New York, NY, USA, 2007, pp. 94–99.

[24] N. Agarwal, H. Liu, L. Tang, P. S. Yu, Identifying the influential bloggers in a community, in: WSDM '08: Proceedings of the international conference on Web search and web data mining, ACM, New York, NY, USA, 2008, pp. 207–218.

[25] G. Mishne, Leave a reply: An analysis of weblog comments.

[26] H. R. Salman Jamali, Diggin digg: Comment mining, popularity prediction, and social network analysis, Tech. Rep. GMU-CS-TR-2009-7, George Mason University (July 2009).

[27] R. Crane, D. Sornette, Robust dynamic classes revealed by measuring the response function of a social system, Proceedings of the National Academy of Sciences 105 (41) (2008) 15649–15653. doi:10.1073/pnas.0803685105.

***Biography of the author(s)**
Click here to download Biography of the author(s): vitae.txt

Jong Gun Lee received his B.S. and M.S. from Kyungpook National University and Korean Advanced Institute of Science and Technology (KAIST), Korea, in 2004 and 2007, in computer engineering and computer science, respectively, and since 2007 he is currently pursuing his Ph.D. degree at at Laboratory of Computer Science (LIP6/CNRS) of University Pierre and Marie Curie (UPMC, Paris 6) in the Networks and Performance Analysis group. His research interests include web contents, online social networks, and network performance measurements.

Sue Moon received her B.S. and M.S. from Seoul National University, Seoul, Korea, in 1988 and 1990, respectively, all in computer engineering. She received a Ph.D. degree in computer science from the University of Massachusetts at Amherst in 2000. From 1999 to 2003, she worked in the IPMON project at Sprint ATL in Burlingame, California. In August of 2003, she joined KAIST and now teaches in Daejeon, Korea. She has served as TPC co-chair for ACM Multimedia and ACM SIGCOMM MobiArch Workshop, general chair for PAM, and TPC for many conferences, including NSDI 2008 and 2010, WWW 2007-2008, COMSNETS 2009, INFOCOM 2004-2006, and IMC 2009. She is currently serving as guest editor for IEEE Network Special Issue on Online Social Networks and Journal of Network and Systems Management Special Issues on New Advances on Meausrement Based Network Management. She won the best paper award in ACM SIGCOMM Internet Measurement Conference 2007 and has been awarded the Amore Pacific Women Scientist Award in 2009. Her research interests are: network performance measurement and analysis, online social networks, and networked systems.

Kave Salamatian is a professor at University of Savoie. His main areas of researches are Internet measurement and modeling and networking information theory. He was previously reader at Lancaster University, UK and associate professor at University Pierre et Marie Curie. KavŽ has graduated in 1998 from Paris SUD-Orsay university where he worked on joint source channel coding applied to multimedia transmission over Internet for his Ph.D.

photo (Jong Gun Lee)
Click here to download high resolution image

photo (Sue Moon)
Click here to download high resolution image

photo (Kave Salamatian)
Click here to download high resolution image