An Approach to Model and Predict the Popularity of Online Contents with Explanatory Factors

Jong Gun Lee, Sue Moon, Kavé Salamatian

An Approach to Model and Predict the Popularity of Online Contents with Explanatory Factors

Jong Gun Lee*, Sue Moon† and Kavé Salamatian‡
*UPMC-Paris Universitas, †KAIST and ‡Université de Savoie
*jonggun.lee@lip6.fr, †sbmoon@kaist.edu and ‡kave.salamatian@univ-savoie.fr

***Abstract*—In this paper, we propose a methodology to predict the popularity of online contents. More precisely, rather than trying to infer the popularity of a content itself, we infer the likelihood that a content will be popular. Our approach is rooted in survival analysis where predicting the precise lifetime of an individual is very hard and almost impossible but predicting the likelihood of one's survival longer than a threshold or another individual is possible. We position ourselves in the standpoint of an external observer who has to infer the popularity of a content only using publicly observable metrics, such as the lifetime of a thread, the number of comments, and the number of views. Our goal is to infer these observable metrics, using a set of explanatory factors, such as the number of comments and the number of links in the first hours after the content publication, which are observable by the external observer.**

We use a Cox proportional hazard regression model that divides the distribution function of the observable popularity metric 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 factors not taken into account. To validate our proposed approach, we use data sets from two different online discussion forums: dpreview.com**, one of the largest online discussion groups providing news and discussion forums about all kinds of digital cameras, and** myspace.com**, one of the representative online social networking services. On these two data sets we model two different popularity metrics, the lifetime of threads and the number of comments, and show that our approach 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.**

I. Introduction

The emergence of Web 2.0 and online social networking services, such as Digg, YouTube, Facebook, and Twitter, has changed how users generate and consume online contents. As the YouTube report of 20 hours worth of video upload every minute demonstrates[1], the amount of user-generated contents is growing fast. Via online social networking services augmented with multimedia contents support, sharing and commenting on other users' contents constitute a significant part of today's Internet users' web experience. Then how do users find contents that are interesting? How do certain contents rise in popularity? If we can predict such rise, we can pick those mostly likely to get popular and filter out others. Such a mechanism will be extremely expedient to users in this age of information deluge.

The popularity of an online content is not a well-defined, but highly subjective term, which can be defined as a mixture of endogenous and exogenous factors. The choice of factors varies from a person to another and from a content to another. Also, we note that accessibility and observability of the data that represent those factors may not be universal. Thus in order to model the popularity of online contents, we first have to decide how we define "popularity". What factors do we take into consideration and which explicit data and related measures shall we use to represent popularity?

Here 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 comments. Individual users have differing views of popularity and measures of choice will be different. Our goal is to develop a general framework that could accommodate differing views by allowing users to choose contributing factors.

Multiple factors, however, complicate the accurate prediction of online contents popularity. The popularity of contents is sometime unpredictable by nature. For example, the flurry of contents and reactions happening very early, even before confirmation, 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 time demonstrating a cascading effect [1], and it is hard to predict what type of contents will eventually instigate such a cascading effect. Last but not least, popularity relates in a complex way to the social psychology of the population of online content users and capturing this intricate relation in a predictive model is difficult. All these difficulties compound the effectiveness of a predictive model.

Szabo and Huberman use a linear regression to predict the long time popularity of an online content from early measurement of user access pattern, based on an observation where the logarithmically transformed popularity of long time popularity of a content is highly correlated with its early measured popularity [2]. For an all-time popular content, their approach, however, produces large error because their purpose is to predict the exact value of its popularity. Our approach in this paper differs from [2] and is rooted in survival analysis. It is used when predicting the precise lifetime of an individual is very hard. A patient with a cancerous metastasis might stay alive much longer than predicted by its doctors, when a healthy young person might die in a car accident. Nevertheless, predicting the likelihood where one will survive longer than a threshold or another individual is possible. In particular one can evaluate the effect of risk factors; smoking is a risk factor that makes a smoker less likely to be alive in a long term

[1] http://www.youtube.com/t/fact_sheet (accessed on Mar 26, 2010)

compared to a non-smoking person. As in survival analysis, we do not aim at knowing the precise popularity of a content but our goal is to infer the likelihood (the probability) that a content with given characteristics will attract popularity above a given threshold.

We use a Cox proportional hazard regression model [3]. It divides the distribution function of the observable 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 factors not taken into account. This approach is frequently applied in biostatistics to model human survival and in reliability theory. The Cox proportional hazard regression model is suitable to our purpose, because the regression does not assume any parametric structure for the baseline hazard, which contains all factors not taken into account. In fact the regression "integrates out", or heuristically removes from consideration, the baseline hazard and maximizes the remaining partial likelihood. The Cox proportional model, therefore, separates the effects of the explanatory parameter from the effects of all other possible factors in the baseline hazard distribution.

We validate our approach over datasets crawled from two online thread-based discussion forums: dpreview.com and myspace.com. Our data sets contain information about 267,000 threads and 2.5 million comments. Defining the popularity of a thread is difficult as these two forums do not provide any information about the statistics about the contents. We assume that the number of comments in a thread and the lifetime of the thread capture the popularity of a thread.

The contributions of our paper are:

1) This work relates the popularity of an online content to explanatory factors (risk factors). We show that survival analysis is applicable to predict the popularity of an online content.
2) We implement the Cox proportional hazard regression model with explanatory variables as risk factors to model and predict a popularity metric.
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 II we explain survival analysis and the Cox proportional hazard regression model and in Section III we describe our prediction methodology. Section IV gives our experiment results of predicting the lifetime of threads and the number of comments of threads. Finally we conclude this paper in Section VI.

II. Background

In this section we give a brief introduction to survival analysis and the Cox proportional hazard regression model.

A. 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 biostatistics and reliability study. Throughout this paper, T represents a random variable denoting the time to a death event and t the wall clock time, respectively. Survival analysis deals with three main functions.

The failure function $F(t)$ is the probability to fail before a certain time t, *i.e.*, the Cumulative Distribution Function (CDF) of the random variable T, $F(t) = Pr\{T \leq t\}$. This definition can be extended to $F(k)$ where k is a discrete increasing variable, such as the number comments on a thread.

The survival function $S(t)$ is the Complementary Cumulative Distribution Function (CCDF) of T, *i.e.*, the dual of $F(t)$, $S(t) = 1 - F(t) = Pr\{T > t\}$. It is the probability of survival up to a certain time 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,

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} \tag{1}$$

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 is related to the survival function through the below relation:

$$H(t) = \int_0^t h(u)\, du = -logS(t) \tag{2}$$

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

$$h(k) = \frac{f(k)}{S(k)} = \left(1 - \frac{S(k-1)}{S(k)}\right) \tag{3}$$

B. Cox Proportional Hazard Regression Model

Cox proportional hazard regression [3] is a semi-parametric approach widely used in practice. In the forthcoming, we describe it given that the failure time is continuous. However, the analysis can be extended in a straightforward way to the case where failure time is a discrete variable k.

The Cox proportional hazard regression fits a regression model defined as a parametric linear function of a set of risk factors to an empirical failure function; it assumes that the hazard function can represented as

$$h(t) = h_0(t) \times exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k). \tag{4}$$

The hazard contains two components: a parametric part that depends linearly on the risk factors and a non-parametric part defined as baseline hazard $h_0(t)$. In other terms, hazard function $h(t)$ is decomposed into two components:

1) The risk factors $\{x_1, ..., x_k\}$ that are the set of factors that influence the survival duration. As the risk factors are introduced into an exponential function, their effects become proportional, *i.e.*, adding to the risk factor has a multiplicative effect on hazard function. Therefore, the coefficients β_i represents the relative importance of risk factors.
2) The baseline hazard $h_0(t)$ that gives the natural risk, *i.e.*, the hazard when any risk factor is not present. No assumptions is made about the form of $h_0(t)$. Using relations between the cumulative hazard function and the

the survival function, one can define a baseline survival function as $S_0(t) = \exp(-\int_0^t h_0(t)dt)$.

C. Interpretation of fitting results

We explain the interpretation of a risk factor in Cox proportional hazard regression with the following example shown in Figure 1. In the example, let's $S(t)$, presented by the dotted

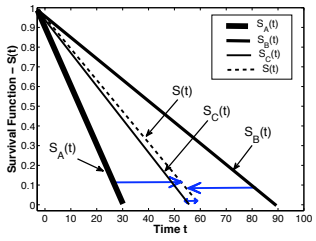


Fig. 1. Examples for understanding risk factors(RFs) and survival function

line, the initial lifetime distribution observed empirically and $S_X(t)$ being the remaining baseline survival function after fitting a Cox model with an unique risk factor $X = \{A, B, C\}$. In order words, $S_X(t)$ is the baseline hazard for the risk factor X, *i.e.*, the survival distribution if the risk factor X is not present, and $S(t)$ is the survival distribution when the risk factor X is present. The wider is the distance between $S(t)$ and $S_X(t)$, the more effective is the risk factor X. To simplify the discussion in this example, we present a survival function as a straight line. Figure 1 implies the followings: Due to the effect of the risk factor A, the overall lifetime is increased. Depending on risk factors, the overall lifetime can be increased, $S_A(t)$, or decreased, $S_B(t)$. Comparatively, risk factor A is more significant risk factor rather than the risk factor C.

III. Proposed Methodology

Formally we model the hazard function $h(t)$ of an observable popularity metric, through a Cox proportional hazard regression, described in Section II. In this context, the popularity metric can be any measure of the popularity of online contents (*e.g.*, the thread lifetime or the number of received comments of a thread in a discussion forum) and the risk factor are chosen to impact the popularity (*e.g.*, the number of comments, the number of contributors, and so on.)

A. Selecting a set of significant risk factors

Similarly to [2], we use as explanatory (risk) factors, early values of the content attributes that are visible to an external user, such as the initial number of comments and the initial number of view.

First of all, one should avoid to use highly correlated factors as they are redundant and can reduce the quality of the fitting. So the first step for selecting the set of significant risk factors is to check the correlation among the potential risk factors in order to rule out the simultaneous usage of highly correlated (the ones with correlation higher than 0.8) factors.

After checking that the risk factors are not highly correlated, we can follow the approach explained in Section II-C to rank different combination risk factors by comparing the baseline survival obtained after fitting the Cox regression model. The further is the resulting baseline survival function from the empirical lifetime distribution, the more significant is this risk factor.

B. Fitting the Cox proportional regression

After setting the set of risk factor to be studied, one can apply the Cox proportional regression to it and try to fit the long-term empirical final distribution of popularities that is obtained after the last activity (comment or view) of a content. However we will fit different regression that are generated using different values of initial observation period. The aim of this step is to find an observation time when the information from risk factors are enough to obtain a good prediction of the likelihoods of popular contents. The quality of the prediction model is assessed by observing the resulting baseline hazard; the further the baseline hazard goes from the empirical distribution the better becomes the predictive power of the variables relative to this observation period.

C. Finding a baseline hazard function for the risk factors

Up to now, we did not provide any functional form for the baseline survival function. The Cox Proportional Hazard regression gives a non parametric description of the baseline survival function that can be thereafter fitted to a parametric distribution. Frequently a Weibull distribution [4] is used. The Weibull distribution is characterized by two parameters: a scale parameter λ and a shape parameter γ:

$$f(x : \lambda, \gamma) = \frac{\gamma}{\lambda}(\frac{x}{\lambda})^{(\gamma-1)} e^{-(\frac{x}{\lambda})^{\gamma}} \tag{5}$$

The CDF of a Weibull distribution is given by Equation 6.

$$Pr(T > t) = 1 - e^{-(\frac{t}{\lambda})^{\gamma}} \tag{6}$$

From Equation (7), we can present a baseline cumulative hazard function like Equation (8).

$$S_0(t) = e^{-(\frac{t}{\lambda})^{\gamma}} \tag{7}$$

$$H_0(t) = (\frac{t}{\lambda})^{\gamma} \tag{8}$$

So $h_0(t)$ can be approximated by $\hat{h}_0(t) = \frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1}$.

D. Forecasting the lifetime likelihood

Having the fitted values of the Cox regression parameters (the β_i obtained in second step) and the parameter of Weibull distribution obtained above, one can retrieve an approximation of the total hazard function through

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

and using Eq. 2 retrieves an approximation of the empirical survival distribution.

IV. Experiment

In this section, we describe our datasets (Section IV-A) and present the experiment result on modeling two different popularity metrics, the lifetime of threads (Section IV-B) and the number of comments per thread (Section IV-C).

A. Datasets

We made two datasets, D-dpreview and D-myspace, from online discussion forum services of forums.dpreview.com and forum.myspace.com and we present the brief description of the datasets in Table I. Dataset D-dpreview contains the

Dataset	Service	Topic	Start - End
D-dpreview	forum.dpreview.com	Canon 40D-10D	2003/01 $\sim$ 2007/12
D-myspace	forums.myspace.com	Music - General	2004/01 $\sim$ 2008/04

dataset	# threads (T)	# comments (C)	U_{all}	U_T	U_C
D-dpreview	140,524	1,496,808	44,955	27,989	41,269
D-myspace	127,607	1,038,989	-	-	-

U_{all}: the number of unique posters
U_T: the num. of unique thread posters, U_C: the num. of unique comment posters

TABLE I
DESCRIPTION OF DATA SETS FOR EXPERIMENT

information of the entire threads and comments of Canon EOS 40D-10D discussion forum for 5 years from 2003 to 2007. We made D-myspace by crawling all of threads and comments of Music-General forum from the creation of the forum to May 2008. For each post, a thread or a comment, we collected its posted timestamp and when making D-dpreview we additionally crawled the anonymized poster identifier for each post[2]. Overall two datasets have more than 267,000 threads and 2.5 millions of comments and in D-dpreview there are about 45,000 unique posters.

Before applying our approach to model the lifetime of threads and the number of comments of threads, we need to define the lifetime of a thread. It can be defined in many ways. We will use the following definition: the lifetime of a thread is the time difference between the posting time of the thread and the posting time of its last comment. However it is not possible to decide if a comment is the definitive last comment. We have therefore to define the lifetime of a thread by assuming that a thread being dead if it does not receive any new comment during a thread expiration time. To decide the expiration time we use inter-comments time and set this value to five days (resp. two days) for D-dpreview (resp. D-myspace) as only 0.5% of comments arrive later[3].

B. Modeling the Lifetime of Threads

In order to model the lifetime of threads of discussion forums, we use the information from user comments as explanatory factors because as an external observer, we can access to these information. From the information, we extract two kinds of information, a) overall user interests on a discussion thread and b) temporal user interests on it. About the overall user interests on a thread, we assume that a user leaves comments on a thread that is interesting to her. Based on this assumption, we consider the following three potential explanatory factors. The last two factors are used to separate the effects of author herself from others.

1. the overall number of comments of a thread
2. the number of comments by author of its original post
3. the number of unique posters except author of the original one

About temporal user interests, we used the inter-commenting times. To receive a high rate of comments for a thread can be a sign when the content is interesting. Thus we consider the following four information as potential factors.

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

Among the above seven potential explanatory factors, we rule out useless factor which do not capture For this, in Figure 2 we plot the empirical survival function of D-dpreview and seven baseline survival functions with seven factors. The line

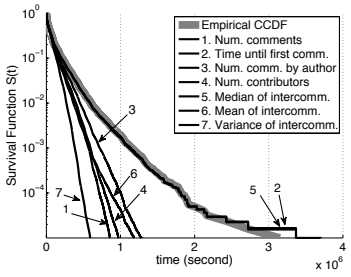


Fig. 2. Selecting significant risk factors (D-dpreview)

named as Empirical CCDF shows the survival function, *i.e.*, the CCDF of the empirical lifetime distribution, obtained over D-dpreview and each curve tagged by a number between 1 and 7 is the baseline survival function when the explanatory factor of the tagged number is used as a risk factor for a Cox regression model. This figure shows that out of seven, two distributions tagged by 2 and 5 are almost same as Empirical CCDF. So we do not use these factors as risk factors and in the following we consider the five potential explanatory factors as below.

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 excluding redundant factors among the five factors, we use correlation coefficient. Correlation coefficient $R(x, y)$ between two variables, x and y, is defined as:

$$R(x,y) = \frac{covariance(x,y)}{\sqrt{variance(x) \times variance(y)}}, \qquad (9)$$

where $|R(x, y)| \sim 1$ means that they are highly correlated but $|R(x, y)| \sim 1$ implies that they are lowly correlated.

Table II shows the correlation coefficient between two factors and it implies that the first explanatory factor is highly

[2]We could not collect any poster identifier for D-myspace because the information was hidden.

[3]Indeed other expiration time can be used such as 99% or 99.9%.

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 II
CORRELATION COEFFICIENT BETWEEN TWO RISK FACTORS (RFS). EACH NUMBER FOR RF IS THE SAME ONE USED IN FIGURE 2

correlated to the fourth one. Thus, instead of using both factors, we use only the first one because it captures much more information than the fourth one as illustrated in Figure 2. Now with four explanatory factors, named with 1, 3, 6, and 7, we make all possible combinations of the factors and then present baseline survival distributions when each combination of factors is used as risk factors in Figure 3. Since based

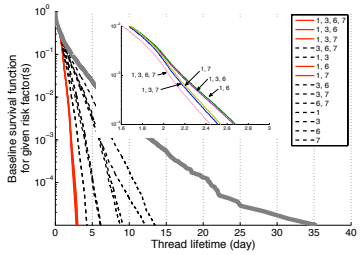


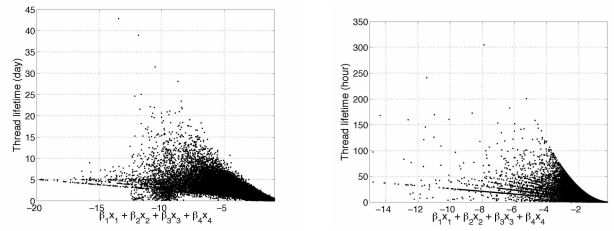
Fig. 3. Ranking the different combinations of risk factors
The rightest line shows the empirical thread lifetime.

on the figure to use all four factors makes the best baseline survival function among them, we use the four factors in our modeling. For D-myspace, we use three factors by excluding 3. the number of comments by a thread poster because we do not have user identifier information in D-myspace.

In Figure 4(a) and 4(d), we plot $S(t)$ and $S_0(t)$ when we introduce the risk factors to Cox proportional hazard models with D-dpreview and D-myspace, respectively. In order to present $S_0(t)$ as a functional form, we fit a Weibull distribution for each baseline hazard function and we present the fitted Weibull distribution for each baseline failure function, $1 - S_0(t)$, in Figure 4(b) and 4(e). From these fittings we can represent a cumulative baseline hazard function with scale and shape parameters of the Weibull distribution and finally we find that $H_0(t)$ are $(\frac{t}{0.4286})^{0.9909}$ for D-dpreview and $(\frac{t}{0.101})^{0.8616}$ for D-myspace.

In Figure 4(c) and Figure 4(f) we show the process to find the minimum observation time for D-dpreview and D-myspace. In each figure, we plot the empirical lifetime distribution (empirical CCDF), the baseline survival function captured during the whole thread duration, and a set of baseline survival functions captured a certain observation time. Remind that the gap between an empirical CCDF and a baseline survival function during the whole thread duration is the information based on the whole thread lifetime. Thus the aim of this step is to find a certain time point where the information captured from the time is close to the information captured from the whole lifetime. For instance, in Figure 4(c) 'BSF after 1 day' shows the baseline survival function when user comment information captured during the first day is introduced to risk factors of a Cox proportional hazard model. The curve of 'BSF after 1 day' is too close to the curve of 'empirical CCDF'. It means that the information captured during the first one day is not enough to predict the lifetime of threads of D-dpreview. Thus based on these figures, we say that we are able to closely predict the empirical lifetime of D-dpreivew and D-myspace by observing the first five to six days and 24 hours, respectively.

By the nature of survival analysis, long-lived instances likely have less hazard than short-lived instances. To see our models follows this fact, we plot risk factor component $(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)$ against the lifetime of threads in Figure 5. Based on this figure we verify that our two models for two datasets produce relatively less hazard values for long-lived threads and relatively much hazard values for short-lived threads[4].



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

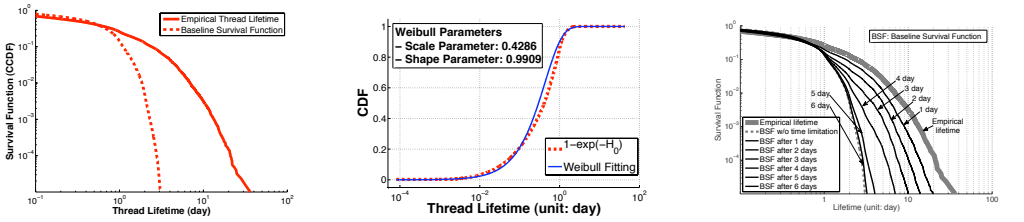
Fig. 5. risk factor component $(x_1\beta_1 ... x_k\beta_k)$ *vs.* the lifetime of threads

To see the relation between an observation time and the risk factor component generated from our models, we show risk factor component against the lifetime of threads varying observation time in Figure 6. In Figure 6(a), we plot the lifetime of threads against their risk factor component with the information of explanatory factors captured during the first day and we could not see clear correlation between them. It means the information during the first day is not enough to predict the popularity metric with our models. As the observation time, however, is getting longer, we can see that our model calibrates the correlation between them and predict the popularity of thread lifetime. In Figure 6(e) about D-myspace, we do not see the correlation between thread lifetime and hazard values, which comes from three-hour observation, but after 24 hours of the creation of threads we can see that the correlation between the lifetime of threads and hazard values with the model from our approach.

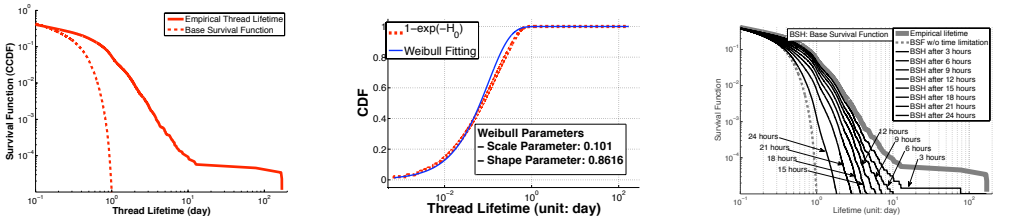
C. Modeling the Number of Comments

We apply our approach to model and predict the number of comments of threads from D-dpreview with the same seven potential explanatory factors used in the previous section. First, to rule out useless factors when modeling the number of comments, we plot the empirical distribution of the number of comments per thread and seven baseline survival functions

[4]We use the word 'relatively' in this sentence because our approach aims to model and predict the likelihood of an objective metric of popularity, not to predict an exact value of the objective metric.

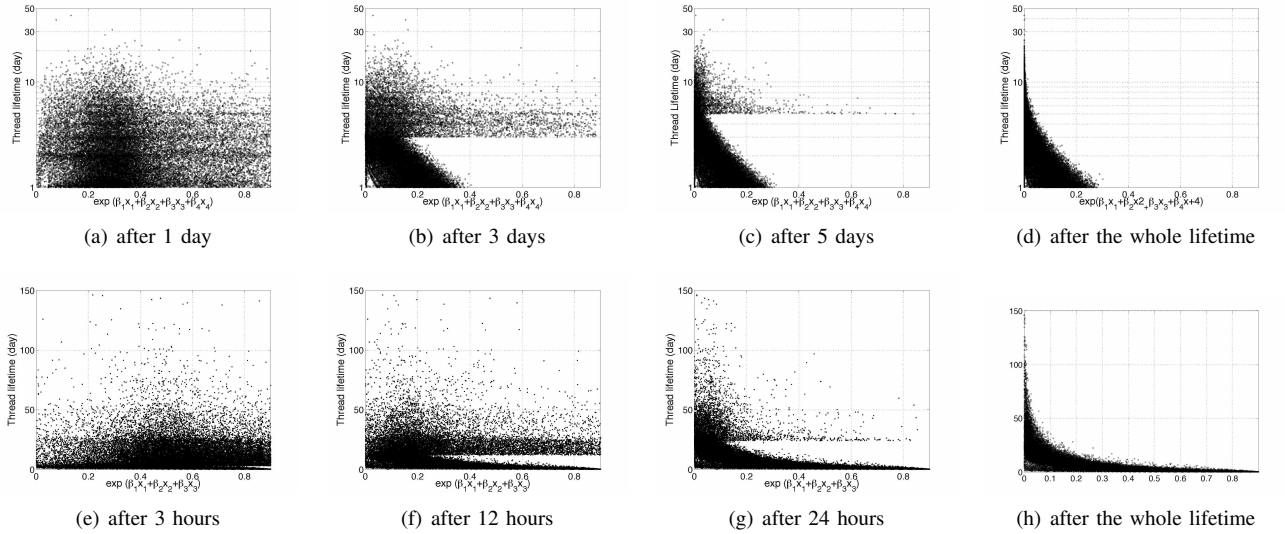


(a) Finding a baseline survival function (b) Fitting it with a Weibull dist. (c) Predicting the lifetime of threads



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

Fig. 4. Prediction of the lifetime of threads from D-dpreview (upper figures) and D-myspace (lower figures)



(a) after 1 day (b) after 3 days (c) after 5 days (d) after the whole lifetime

(e) after 3 hours (f) after 12 hours (g) after 24 hours (h) after the whole lifetime

Fig. 6. Risk factor component *vs.* Thread lifetime, varying observation time.(The upper three figures comes from D-dpreview and the bottom figures comes from D-myspace.)

with seven factors in Figure 7(a). Based on this figure, we choose the following three factors: the number of comments, the number of comments by a thread poster, and the number of unique posters. Then we check whether any two factors are highly correlated with Table II. Since any two factors are not highly correlated, we now make all possible different combinations with the factors. With the combinations, in Figure 8, we plot baseline survival functions when each combination of factors is used for risk factors. This figure shows that three combinations of factors illustrated by straight lines capture more than other ones presented by dotted lines. Amongst three combinations, we use all three explanatory factors as the risk factors for our model to predict the number of comments.

With these factors, we compute their baseline survival function and present it as well as the empirical survival function in Figure 7(b). To provide the baseline survival function as a functional form, we fit it with a Weibull distribution and show

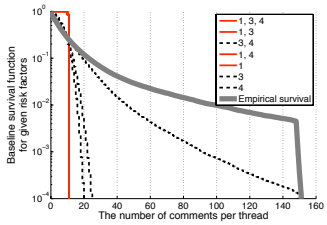
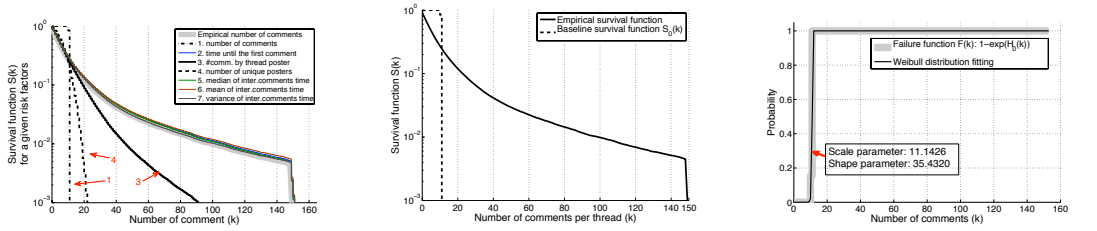


Fig. 8. Ranking the different combinations of risk factors

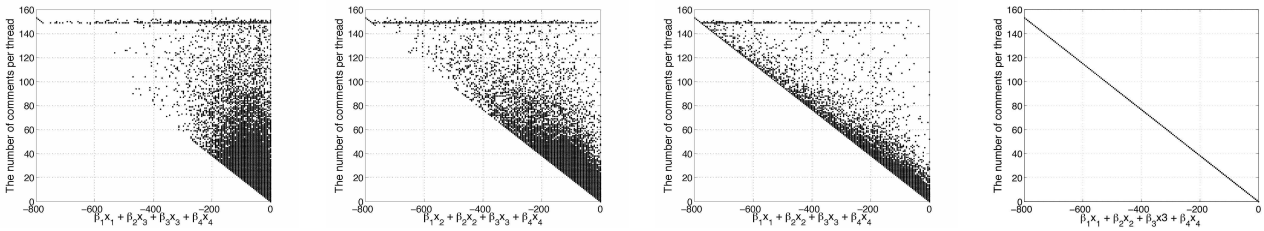
the fitting result in Figure 7(c). Since the scale and shape parameters of the fitted Weibull distribution are 7.4189 and 1.8496, respectively, the cumulative baseline hazard function $H_0(t)$ is determined as $(\frac{t}{7.4189})^{1.8496}$.

Now we find the minimum observation time to predict the number of comments per thread. For this, we vary observation



(a) Selecting major risk factors (b) Baseline survival function (c) Fitting it with a Weibull distribution

Fig. 7. Predicting the number of comments of online discussion forum threads



(a) after 12 hours (b) after 1 day (c) after 3 days (d) after the whole lifetime

Fig. 10. Risk factor component *vs.* Number of comments per thread (varying observation time)

time as shown in Figure 9. This figure implies that when

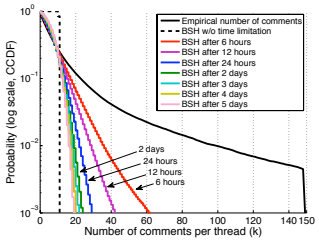


Fig. 9. Determining a minimum observation time

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.

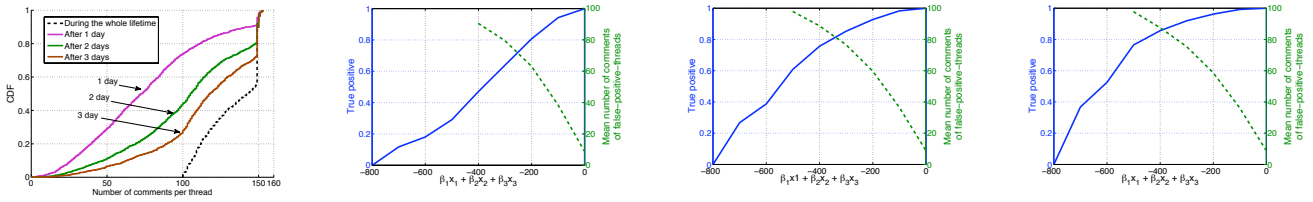
In Figure 10, we show risk factor component *vs.* the number of comments per thread, varying observation time. Especially, Figure 10(d) shows the correlation between a hazard value and the number of comments of a thread when all thread are dead. It clearly implies that while a thread with much hazard has less comments, a thread with less hazard has much comments. In other words, as the values of risk factors of a thread are increasing, the hazard of the thread is decreasing and its number of comments is increasing. Now let us visit Figure 10(a), 10(b), and 10(c). In Figure 10(a), we see that hazard values of almost all thread are high by positioning at the right part of x-axis, but the hazard values are decreasing as the observation duration is getting longer in Figure 10(b) and 10(c).

Now we bring an application to predict threads, each of which has more than 100 comments. We find that there are 1,406 threads which have received more than 100 comments in D-dpreview and in Figure 11(a) we plot how many comments they received after one, two, and three days. After one day (two and three days) about 24% (56% and 73%) threads among 1,406 have more than 100 comments (respectively.) The following three figures show that how accurately we can predict them and what the mean of comments of mis-predicted threads after one, two, and three days. For instance, when we choose -200 as a threshold of risk factor component, we can correctly find about 80% of threads after one day, based on Figure 11(b). We additionally have mis-predictied threads ('false positive' threads), which the mean of their comments is about 63. In a similar way, we can identify about 80% of correct threads after two days based on Figure 11(c) when taking -355 of a threshold. Remind that the mis-predicted threads by our model are somehow popular even though they are less popular than correctly identified ones, because our approach is based on the likelihood of the objective metric. Thus one who adopts our approach to model and predict an objective metric of the popularity of a kind of online contents can choose her threshold which satisfies her aim of prediction in terms of a precision and required observation time.

V. Related Work

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

- *Survival analysis*
  Survival analysis [5] has been applied to various areas, such as bio-medical science, sociology, and epidemics [6], [7], [8], [9]. Among the methodologies for survival analysis, Cox proportional hazard regression model [3], which is a semi-parametric survival analysis methodology, has been widely used [10], [11], [12]. In this paper, we first adopted survival analysis and Cox proportional



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

Fig. 11. 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.)

regression approach to model and predict the popularity of online contents.

- *Analysis on Threads and Comments*
  The authors of [13], [14] analyzed the posts and comments of Slashdot. In detail, the authors of [13] explained the behaviors of inter-posts times with statistical models and the authors of [14] focused on to analyze the dynamics of posts and users. There was a macroscopic-level analysis result, such as analysis on the average views and incoming links about posts and comments with web logs in [15]. In [16], [17] the information of user comments was used to understand user intention, and in [18] to find influential authors based on user comments was investigated.
- *Modeling Inter-Posting or Predicting Popularity*
  In [13], authors modeled post-comment-interval with four different statistical models and they predicted intermediate and long-term user activities. [2] proposed a methodology to predict the popularity of online contents based on a finding, the correlation of popularity between early and later times. Then the authors proposed three prediction models and validated them with Youtube and Digg datasets. In [19], authors built a co-participation network among Digg users with comment information of their Digg dataset and proposed a method to predict the popularity of online using an entropy measure explaining user interest peak and the co-participation network. Our work is different in a point that we model and predict the popularity of online contents with a set of explanatory factors by applying survival analysis and Cox proportional hazard model.

VI. Conclusion

In this paper, we proposed a methodology about macroscopic prediction of the popularity of online contents, which is to infer the likelihood that a content will attract a popularity. To model and predict an objective metric of the popularity of online contents we apply Cox proportional hazard regression model for a set of given explanatory factors. We validated our approach by predicting two kinds of popularity features (thread lifetime and the number of comments per thread) with two datasets from two discussion online forums. In the experiments, we showed that our approach successfully modeled a popularity metric with a set of risk factors and the popularity metric was determined by the information represented by the risk factors.

References

[1] M. Cha, A. Mislove, and 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*. New York, NY, USA: ACM, 2009, pp. 721–730.

[2] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Social Science Research Network Working Paper Series*, November 2008.

[3] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972.

[4] W. Weibull, "A statistical distribution function of wide applicability," *Journal of Applied Mechanics*, pp. 293–297, 1951.

[5] R. Schlittgen, "Survival analysis: State of the art," *Computational Statistics and Data Analysis*, vol. 20, no. 5, pp. 592–593, November 1995.

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

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

[8] A. Diekmann, M. Jungbauer-Gans, H. Krassnig, and S. Lorenz, "Social status and aggression: a field study analyzed by survival analysis." *J Soc Psychol*, vol. 136, pp. 761–768, Dec 1996.

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

[10] J. Heckman and B. Singer, "The identifiability of the proportional hazard model," *Review of Economic Studies*, vol. 51, no. 2, pp. 231–41, April 1984.

[11] P. B. Seetharaman and P. K. Chintagunta, "The proportional hazard model for purchase timing: A comparison of alternative specifications," *Journal of Business & Economic Statistics*, vol. 21, no. 3, pp. 368–82, July 2003.

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

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

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

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

[16] M. Hu, A. Sun, and 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*. New York, NY, USA: ACM, 2007, pp. 901–904.

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

[18] N. Agarwal, H. Liu, L. Tang, and 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*. New York, NY, USA: ACM, 2008, pp. 207–218.

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