Are We Still Friends: Kernel Multivariate Survival Analysis

Shiyu Liang†, Ruotian Luo†, Ge Chen†, Songjun Ma†, Weijie Wu‡, Li Song†, Xiaohua Tian†, Xinbing Wang†,
†Dept. of Electronic Engineering, Shanghai Jiao Tong University, China
‡School of Information Security Engineering, Shanghai Jiao Tong University, China
Email:{lsy18602808513, skylikelrt, chenge, masongjun, weijiewu, song_li, xtian, xwang8}@sjtu.edu.cn

***Abstract*—Online Social Network becomes the most prevalent platform for exchanging information between users, maintaining friendships online. As is well-known to us, however, some friendships even those intimate ones might vanish. Therefore, precisely modeling and predicting state of each online relationship is worthwhile in many respects. For social communication services such modeling permits new and novel online services. In addition, constructing this model might enlighten us in exploiting information spreading pattern in online social network. In this paper, we propose a model in determining a probability distribution which describes the 'surviving time' of each friendships by applying one commonly used method in sociology, survival analysis. We discuss a series of social explanatory variables that highly affect this probability distribution. Moreover, methods in the moving average process are devoted to determining the appropriate parameter in survival model. Furthermore, to avoid the high computational complexity in kernel learning we impose sparsity in our model. Finally, with the experiments on real data, the proposed survival model is proven to be of high accuracy, and thus of great potential for further applications.**

I. Introduction

Thanks to the last generation of mobile devices offering efficient ways to generate and share information among the electronic world, online social network now becomes the most advanced and prevalent means of communication at our disposal to socialize with friends. However, a phenomenon seems to be familiar to all online users that some acquaintances whom we know little about and barely have communication with, in some sense, are not friends. At the same time, some friends, even those intimate ones will gradually alienate themselves from us. Thus, a very intriguing question is whether these acquaintances are our friends, or in other ways, what is the probability distribution that the existed friendship vanishes with respect to time.

Specifically evaluating the current state of dynamic friendships is necessary not only in studying the online social network structure, but also in both understanding the spreading pattern of online information and further designing social network systems [1]. As a concrete example, by precisely predicting the strength of online friendships, online service providers such as Facebook, Twitter and LinkedIn could provide novel services or modify their existed online advertising model to avoid losing subscribers or to make profits more efficiently. Furthermore, enterprises could maintain their influence in industries in order to hold market share as well as to enhance their status in business circles. More importantly, in U.S. elections, both camps yearn for enhancing their influence to obtain more votes. Consequently, numerous studies concerned with political science have been proposed so that the immediate and extensive property of online social network could be used for political propaganda. Among these studies, one remarkable trial concerning with election messages spread during the Congressional midterms in 2010, consisted of 61 million Facebook users, show that the messages magnificently influenced the users and their friends and that transmission almost occurred between 'close friends' [2]. At the same time, as Dunbar, a British anthropologist, and a series of recent researchers pointed out, the offline [3] as well as the online [4] network are both formed of relationships having different social characters such as intimacy and contact frequency. Therefore, to be able to design online services for numerous purposes and understand the online messages spreading pattern more specifically, we must determine the probability that two friends will become just acquaintances and the social characters that highly influence this probability distribution.

In this study, we assume that the information could be merely spread through those 'living' friendship, which means that we consider that the possibility of messages exchanging through those 'dead' relationships, defined as contacts with those acquaintances with whom we barely have communication, is infinitesimal. In addition, the state of each friendships, 'living' or 'dead', is changeable with the time lapsing. In this perspective, the online friendships have different 'surviving time', controlled by the social properties of the two individuals constructing it.

It is, however, surprising to find that a commonly used analysis in medical studies, the survival analysis, is extremely adequate for analyzing the probability distribution of 'surviving time' for each friendship. The survival analysis is a field of traditional statics concerned with time before the occurrence of some special event, for instance, death. Many applications have appeared in the sociology studies, such as using the hazard function to evaluate the extent at which each social characters affect the friendship evolution [5] and modeling the popularity of online messages [6] by incorporating a series of social explanatory characters in to survival analysis. In most cases, however, the collected data is sporadically discrete, which means each events we observe occurs incidentally and therefore the first essential difficulty in applying the survival analysis is finding a suitable smoothing method so that we

could construct a dynamic model. The smoothing process is usually done by applying the moving average process but they merely coarsely estimate the smoothing length [5] or directly set the smoothing length equal to the total observing duration [4]. Although these two ways seem practical and efficient, it is indeed inappropriate for constructing a systematic model with high accuracy. At same time, due to the uncertainty of dependencies between the social characters and surviving time, the second difficulty manifests itself in the choice of learning algorithm. Although a support vector machine learning method with kernel function was proposed to fit in nonlinear dependencies, a learning algorithm with computational complexity up to $O(n^3)$ and memory requirement up to $O(n^2)$ is equally unacceptable especially considering the enormous size of training data set.

The aim of this paper is to construct a dynamic model so that the 'living' probability distribution of friendship could be precisely described through social characters by applying the multivariate survival analysis. Especially, we analyzed a Sina Microblog data set, containing communication records of more than 40,601 accounts, seeking whether it is possible to determine the parameters used in multivariate survival model. We first proposed a method of choosing the smoothing order in moving average process so that the amount of friendships could be calculated dynamically with respect to time. Second, we constructed a multivariate survival model in this paper to depict the probability distribution of each friendships and incorporate methodology from kernel learning so that, theoretically, the constructed survival model is capable of learning arbitrary dependencies between social characters and surviving time. In addition, in order to lower down the computational complexity and storage requirement in kernel learning algorithm, we imposed the sparsity in the kernel matrix. Finally, we evaluated the accuracy of our proposed survival model.

The paper is organized as follows: in Section II, we introduce the underlying properties and commonly used methods in moving-average model to preprocess data. Then, in Section III, we introduce the basic model used to determine the probability distribution of surviving time of each friendships with various social explanatory variables. In Section IV, we present our results, discussing their meaning and showing the performance of the proposed multivariate survival model. Finally, in Section V we draw the main conclusion of this analysis.

II. Data Preprocessing

The dynamic social friendships produce observable 'spikes' of online communication [5]; therefore it is possible to create a smooth approximation of the instantaneous network by applying the moving average model with an order of m. By applying moving-average process, we thus consider that, at time t, an online social friendship is absent in the immediate network if and only if there is no information exchanged in both directions during $(t-\tau, t]$, where the smoothing window length τ is given by $\tau = m + 1$. In this section, we aim to introduce an approach to determining the process order in moving-average model.

The moving-average model is commonly encountered in approaches of modeling univariate time series model [7] and the order of the moving-average, to a large extent, determines the model performance. For a concrete example, when the smoothing window length τ is chosen too short, it will be easy to find that some friendships will reconstruct after having been terminated. On the contrary, when τ is chosen too long, some remote communications which seems unlikely to have impacts on the current structure of the online social network will nevertheless be included in the calculation of social friendship strength. As a result, the necessity of determining the order m or the smoothing length τ aptly such that the survival model could be well-performed is self-evident.

Considering the observation that the average vertex degree is in equilibrium with respect to time t [5], consequently we suppose that $y(t)$ is the friendships increase within $(t-1, t]$ and $\epsilon(t)$ is the unknown added white noise. Therefore based on our moving-average model, the increase $y(t)$ could be represented by

$$y(t) = \beta_0\epsilon(t) + \beta_1\epsilon(t-1) + ... + \beta_m\epsilon(t-m) \quad (1)$$

and the expectation of white noise is zero

$$E[\epsilon(t)] = 0, \quad (2)$$

where the order m and the parameters β_i are assumed to be unknown. In addition, considering the general property [8] of a moving-average process

$$E[y(t)y(t-i-1)] = 0, \quad for\ i \geq m, \quad (3)$$

therefore (3) could be used as a test for determing m. Considering the statistic property of

$$\begin{aligned}\sigma(m+j) &= E[y(t)y(t-(m+j))] \\ &= \frac{1}{N-m-j+1}\sum_{t=m+j}^{N} y(t)y(t-m-j), \quad (4)\end{aligned}$$

where j is an arbitrary positive integer and N is the length of time series and noticing the property that $\sigma(m+i)$ and $\sigma(m+j)$ are independent of each other when $|i-j| > m$ satisfies, thus the testing condition could be replaced by

$$E[y(t)y(t-i-1)] = 0, \quad for \quad i = m, ..., 2m-1. \quad (5)$$

As a result, the random variables $\sigma(m+1), \sigma(m+2), ..., \sigma(2m)$ then could be used to form the hypothesis test for determining the order.

Due to the uncertainty of the distribution of white noise process, the joint probability density distribution function of $\sigma(m+1), \sigma(m+2), ..., \sigma(2m)$ could not be represented explicitly. However, noticing N is sufficiently large, the joint probability density could be approximated by

$$f(\sigma(m+1), ..., \sigma(2m)) = \frac{1}{\sqrt{(2\pi)^m\,|\boldsymbol{\Psi}|}} e^{\boldsymbol{\sigma^T \Psi^{-1} \sigma}}, \quad (6)$$

where the elements of $\mathbf{\Psi}$ are given by

$$\psi_{ij} = \begin{cases} \frac{\sigma^2(0)}{N} + \frac{2}{N}\sum_{k=1}^{m+i}\sigma^2(k), & \text{for } i = j \\ \frac{2}{N}\sum_{r=0}^{m+j}\sigma(r)\sigma(r+i-j), & \text{for } i > j \\ \frac{2}{N}\sum_{r=0}^{m+i}\sigma(r)\sigma(r+j-i), & \text{for } i < j \end{cases} \tag{7}$$

and the vector of random variables is represented by

$$\boldsymbol{\sigma^T} = [\sigma(m+1) \quad \sigma(m+2) \quad ... \quad \sigma(2m)]. \tag{8}$$

In our hypothesis testing, we will reject the null hypothesis (order is m) with probability of type-1 error equal to θ if $\left|\boldsymbol{\sigma^T}\mathbf{\Psi^{-1}}\boldsymbol{\sigma}\right| \geq l^2$; l is chosen so that

$$\frac{1}{(2\pi)^{m/2}}\int_0^l exp(-\frac{r^2}{2})S(r)dr \tag{9}$$

where $S(r)dr$ is the spherically symmetric volume in a m-dimensional space. Find l^2 so that the probability is 0.95; then we will accept the null hypothesis if the sample of $\boldsymbol{\sigma}$ satisfies $\left|\boldsymbol{\sigma^T}\mathbf{\Psi^{-1}}\boldsymbol{\sigma}\right| \leq l^2$.

In summary, each time we construct a hypothesis test, we could determine whether the order we suppose currently should be considered as an appropriate value. After constructing a series of hypothesis tests, we could consequently determine a region where all the value are adequate candidates for the order m.

III. System Model

After having smoothing the discrete data, we then could construct the kernel multivariate survival model. In this section, we consider a homogeneous population of online friendships, each having a 'surviving time'. The first part of this section contributes to a brief summarization of multivariate survival model, full introduction included in [9]. In the second part, we then introduce the approaches to survival analysis which uses sparse kernel learning methods to efficiently calculate the optimal parameters.

A. Multivariate Survival Model

For a given statistical distribution describing time to death T

$$F(t) = P(T \geq t), \tag{10}$$

the survival analysis aims to determine its optimal parameters. Noticing that this differs from the conventional form of the cumulative distribution function, therefore the probability distribution function is given by

$$f(t) = -F'(t). \tag{11}$$

Meanwhile, another essential definition in survival model called as hazard function is defined as

$$h(t) = \frac{f(t)}{F(t)}, \tag{12}$$

which enlightens us to consider the immediate 'risk' attaching to a friendship known to be alive at time t. Actually, there are huge amount of nonnegative distributions which could be used for survival model. In this paper, we adopt one of the commonly used distribution, the exponential distribution, of which the cumulative distribution function, the probability distribution function as well as the hazard function is calculated by

$$F(t) = e^{-\rho t}, \quad f(t) = \rho e^{-\rho t}, \quad h(t) = \rho \tag{13}$$

respectively, where the variable ρ controls the time scale of this exponential distribution and is determined by the social property of each friendship. For a given friendship, the constant hazard coefficient reflects the property of our survival model reasonably called as lack of memory. The optimal parameters are determined according to the maximum likelihood criterion. For a given training data set $D = \{t_i\}_{i=1}^{l}$, keeping track of surviving time of l friendships, without loss of generality, we assume that the data are the samples of an independent and identical distribution from an unknown distribution. Consequently, the data likelihood is calculated as the product of density function of each observed friendships,

$$P(D) = \prod_{i=1}^{l} f(t_i). \tag{14}$$

Then the optimal parameters could be determined by minimizing the negative logarithm of this likelihood function.

However, a special source of difficulty in feeding model with data is the probability that some friendships may not be observed for the full time from birth to death. Therefore, in survival analysis, the friendships where either born or extinct time are not observed are said to have been 'censored'. Clearly, although the born or extinct time is not given, the censored data should also be considered in training the survival model, since they also bring qualified though limit information of the online relationships. Adopting some simple tricks, the censored data could as well be merged into our previous likelihood function and thus the likelihood function becomes

$$P(D) = \prod_{i\in U} f(t_i) \prod_{i\in C} F(t_i), \tag{15}$$

where the index set of the censored and uncensored data are represented by C and U individually.

Returning to our online social network model, it is expected that the two individuals forming a valid friendship do have some social properties which could be represented by quantified variables namely common acquaintances, strong indirect connections and gender. Hence in our application, we attempt to construct a model to determine the distribution of surviving time through the explanatory variable vector set $\left\{\boldsymbol{x_i} \in X \subset R^d\right\}_{i=1}^{l}$, where each element $\boldsymbol{x_i}$ is composed of d variables, representing one combination of social properties. In this paper, the following variables considered:

1) Strong indirect relationship: for every two users the

indirect interaction is computed as

$$\omega_{ij}(t) = \frac{1}{k_{ij}\tau} \sum_{q=1}^{k_{ij}} \sqrt{(m_{iq} + m_{qi})^2 + (m_{jq} + m_{qj})^2}, \tag{16}$$

where k_{ij} is the number of mutual contact friends possessed by user i and j, and m_{iq} is the amount of messages transmitted from user i to q observed within $(t-\tau, t]$. The sum $m_{iq} + m_{qi}$ is therefore the total amount of messages transmitted between user i and q within $(t-\tau, t]$. Consequently, the indirect relationship strength is calculated as the time average of indirect interactions between each pair $\omega_{ij} = \frac{1}{N}\sum_{t=1}^{N} \omega_{ij}(t)$.

2) Mutual acquaintances: the number of mutual acquaintances, at the time of sampling, which does not mean these acquaintances should communicate with either members of pairs.
3) Gender: 0 if genders of both users are same, 1 otherwise.

Apparently, all the variables illustrated above more or less have impacts on the friendship formation or extinction and thus we will incorporate these multivariate parameters into previous survival model. However, noticing that the scale of each explanatory variable varies severely, these variables should then be normalized into variables with mean of 0 and variance of 1 before applying the learning algorithm. As a result, instead of being controlled merely by the constant coefficient ρ, the scale of the distribution function is determined by a function $\rho(\boldsymbol{x})$ according to the values of the variables. Substituting the variable function $\rho(\boldsymbol{x})$ into (13), the cumulative distribution function then becomes

$$F(t; \boldsymbol{x}) = e^{-\rho(\boldsymbol{x})t} \tag{17}$$

and the density function becomes

$$f(t; \boldsymbol{x}) = \rho(\boldsymbol{x}) e^{-\rho(\boldsymbol{x})t} \tag{18}$$

while the hazard function becomes

$$h(t; \boldsymbol{x}) = \rho(\boldsymbol{x}). \tag{19}$$

Noticing the truth that the probability density function $f(t; \boldsymbol{x})$ is nonnegative for arbitrary $\boldsymbol{x}$ and t, thus the most prevalent form is the exponential expression, given by

$$\rho(\boldsymbol{x}; \boldsymbol{w}, b) = exp\{\boldsymbol{w} \cdot \boldsymbol{x} + b\} \tag{20}$$

where the shape of the baseline function as well as the parameters $(\boldsymbol{w}, b)$ could be determined by maximizing the data likelihood function (15). Although the multivariate model is the most elementary model in survival analysis, it does be appropriate for huge scenarios.

B. Incorporating Kernel Learning Methods

The kernel learning methods aim to construct a fixed mapping function $\phi(\boldsymbol{x})$ to transform the input space X into the feature space F, $\phi(\boldsymbol{x}) : X \to F$, such that the variable function becomes $\rho(\boldsymbol{x}; \boldsymbol{w}, b) = exp\{\boldsymbol{w} \cdot \phi(\boldsymbol{x}) + b\}$. According to the maximum likelihood criterion, the optimal regression coefficients are determined by minimizing the loss function

$$L(w, b) = -ln\left\{\prod_{i \in U} f(t_i) \prod_{i \in C} F(t_i)\right\}, \tag{21}$$

which is given by the negative logarithm form of data likelihood function. Instead of specifying the mapping function $\phi(\cdot)$ explicitly, a kernel function is used to represent the inner product of two mapping functions in the feature space, i.e. $k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{x}')$. According to Mercer Theorem [10], it is necessary and sufficient that the for any $\boldsymbol{x}$ the kernel matrix $K = [k_{ij} = k(x_i, x_j)]_{i,j=1}^{l}$ should be a symmetric positive semidefinite matrix. In this paper, therefore, we adopt the linear form of polynomial kernel function

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x} \cdot \boldsymbol{x}' + 1. \tag{22}$$

In representer theorem [11], the minimizer of the loss function (22) can be written as

$$\boldsymbol{w} = \sum_{i=1}^{l} \alpha_i \phi(\boldsymbol{x_i}) \tag{23}$$

such that that the output of the model is the linear combination of the evaluation of the kernel function

$$\rho(\boldsymbol{x}; \boldsymbol{\alpha}, b) = exp\left\{\sum_{i=1}^{l} \alpha_i k(\boldsymbol{x_i}, \boldsymbol{x}) + b\right\}. \tag{24}$$

Substituting (18), (19), (25) into (22), the loss function then could be written as

$$\begin{aligned} L(\boldsymbol{\alpha}, b) \quad = \quad & \sum_{i=1}^{l} \left[t_i exp\left\{\sum_{j=1}^{l} \alpha_j k(\boldsymbol{x_j}, \boldsymbol{x_i}) + b\right\}\right] \\ & - \sum_{i \in U} \left(\sum_{j=1}^{l} \alpha_j k(\boldsymbol{x_j}, \boldsymbol{x_i}) + b\right). \end{aligned} \tag{25}$$

The kernel survival analysis model, however, is apparently full dense due to the high possibility that almost all coefficients α_i are nonzero. Meanwhile, noticing that our training algorithm has the computational complexity of $O(l^3)$ and the memory requirement of $O(l^2)$, it is thus necessary to approximate our kernel matrix K such that the rank of the approximated matrix is far lower than the size of training set l. According to Fine and Scheinberg [12], we therefore could use the incomplete Cholesky factorization with symmetric permutations to find a symmetric submatrix $\hat{K}$ of K [13] with a full-rank of M. Meanwhile, some other results of this algorithm are equally remarkable that there are M columns of K contribute to constructing $\hat{K}$ and that the rest column of K is almost linearly dependent on these columns and could be deleted before training with slightly affecting model performance. The variable function of the sparse kernel model then could be represented by

$$\rho(\boldsymbol{x}; \boldsymbol{\beta}, b) = exp\left\{\sum_{i=1}^{M} \beta_i \hat{k}(\boldsymbol{x_i}, \boldsymbol{x}) + b\right\}, \tag{26}$$

where $\hat{k}$ is the sparse kernel function and the loss function becomes

$$L(\boldsymbol{\alpha}, b) = \sum_{i=1}^{l}\left[t_i exp\left\{\sum_{j=1}^{M}\beta_j\hat{k}(\boldsymbol{x_j},\boldsymbol{x_i})+b\right\}\right] - \sum_{i\in U}\left(\sum_{j=1}^{M}\beta_j\hat{k}(\boldsymbol{x_j},\boldsymbol{x_i})+b\right). \quad (27)$$

Considering the convex property of the loss function, then the optimal regression coefficients $\boldsymbol{\omega} = (\boldsymbol{\beta}, b)$ could be determined merely using the second-order gradient descent optimization approaches. Let $\boldsymbol{\gamma}$ and $\boldsymbol{H}$ represent the gradient vector and the Hessian matrix of L respectively

$$\boldsymbol{\gamma} = \left(\gamma_r = \frac{\partial L}{\partial \omega_r}\right)_{r=1}^{m}, \quad \boldsymbol{H} = \left[h_{rs} = \frac{\partial^2 L}{\partial\omega_r\partial\omega_s}\right]_{r,s=1}^{l} \quad (28)$$

where

$$\gamma_r = \begin{cases} \sum_{i=1}^{l}\left[t_i\hat{k}(\boldsymbol{x_r},\boldsymbol{x_i})\varphi_i\right] - \sum_{i\in U}\hat{k}(\boldsymbol{x_r},\boldsymbol{x_i}), & \text{for } \omega_r \neq b \\ \sum_{i=1}^{l} t_i\varphi_i - \sum_{i\in U}1, & \text{for } \omega_r = b \end{cases} \quad (29)$$

and

$$h_{rs} = \begin{cases} \sum_{i=1}^{l}\left[t_i\hat{k}(\boldsymbol{x_r},\boldsymbol{x_i})\hat{k}(\boldsymbol{x_s},\boldsymbol{x_i})\varphi_i\right], & \text{for } \omega_r \neq b,\ \omega_s \neq b \\ \sum_{i=1}^{l}\left[t_i\hat{k}(\boldsymbol{x_s},\boldsymbol{x_i})\varphi_i\right], & \text{for } \omega_r = b,\ \omega_s \neq b \\ \sum_{i=1}^{l}\left[t_i\hat{k}(\boldsymbol{x_r},\boldsymbol{x_i})\varphi_i\right], & \text{for } \omega_r \neq b,\ \omega_s = b \\ \sum_{i=1}^{l} t_i\varphi_i, & \text{for } \omega_r = b,\ \omega_s = b \end{cases} \quad (30)$$

while φ_i is the abbreviation of

$$\varphi_i = exp\left\{\sum_{j=1}^{m}\beta_j\hat{k}(\boldsymbol{x_j},\boldsymbol{x_i})+b\right\}. \quad (31)$$

Finally, the model parameters could be iteratively optimized according to the following update formula:

$$\boldsymbol{\omega}^{new} = \boldsymbol{\omega}^{old} - \boldsymbol{H}^{-1}\boldsymbol{\gamma}. \quad (32)$$

In summary, through constructing the survival model with various social explanatory variables, we then could form a model in describing the probability distribution of each friendships, meanwhile the incorporation of learning and sparsity is devoted to optimizing the parameters efficiently and precisely.

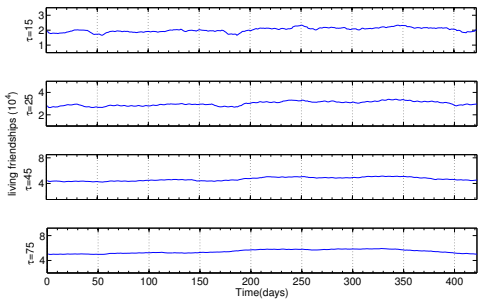

Fig. 1. Living friendships in online social network over time, for four choice of smoothing window τ=15, 25, 45, 75 days.

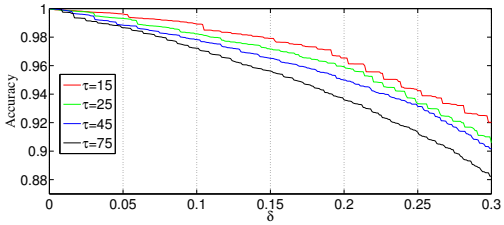

Fig. 2. Accuracy of predicting surviving time over variable δ, for four choice of smoothing window τ=15, 25, 45, 75 days. The accuracy stays highly even smoothing window changes severely.

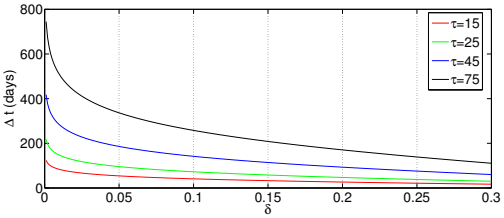

Fig. 3. Average duration Δt of transitional period over variable δ, for four choice of smoothing window τ=15, 25, 45, 75 days. The average duration damps rapidly when δ increase for each smoothing window.

IV. Experimental Evaluation

The data set we use is from the Sina Microblog, containing 40,601 online users, 1,261,304 message transmission and 273,053 online friendships observed from August 7th, 2012 to December 17th, 2013. In this section we provide a experiment to demonstrate the performance of our proposed multivariate survival model. Furthermore, comparisons against different values of smoothing window τ are devoted to illustrating the extent to which the time scale affects the survival model.

Using the data preprocessing methods introduced in section II, we first calculate the appropriate value set $\{\tau|20 \leq \tau \leq 30\}$ from which the adequate smoothing window length is chosen. We then let $\tau = 15, 25, 45, 75$ days. Therefore, we could compare the variation of the number of friendships at time t and summarize the result in Figure 1, from which we could see that friendship number exhibits varying level of stability over time and with respect to smoothing length τ;

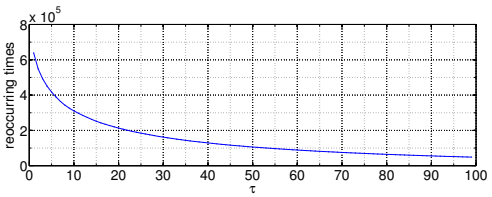

Fig. 4. The amount of reoccurring relationships over variable smoothing window τ. The reoccurring amount decrease rapidly with τ increasing; the fewer times of reoccurrence means the better performance of proposed survival model.

and that a bigger value of τ results in a steadier oscillation in amount of friendships in network. In addition, Figure 1 clearly shows that as the smooth window τ increases, the amount of living friendships increase. As a result, is seems that a wider smoothing window is more suitable for preprocessing data.

We second compare the accuracy of the survival model in terms of predicting time when a existed friendship will vanish. Given a fixed coefficient δ devoted to representing the probability below which we will consider the friendship as dead one, at the same time, we assume that the relationship survives when cumulative probability is larger than $1-\delta$. Consequently, for each friendship in the training set, we then could calculate two time scale t_l and t_h, representing the time of the end of the living period and that of the beginning of the dead period respectively, such that t_l and t_h suffice

$$F(t_l) = 1-\delta, \quad F(t_h) = \delta \tag{33}$$

where $F(t)$ is cumulative probability distribution function in multivariate survival model in Section III. Therefore, the duration of transitional period is computed as $\Delta t = t_h - t_l$ where the state of each friendship is described by cumulative function $F(t)$. Meanwhile, the coefficient δ should suffice $0 \leq \delta \leq 0.5$ due to $F(t_l) \geq F(t_h)$. As a result, the accuracy is calculate by the fraction of friendships whose living time t_i recorded in training set satisfies $t_l \leq t_i \leq t_h$. In this paper, we calculate the accuracy for each δ chosen from 0 to 0.3 and the result is shown in Figure 2, from which we could clearly find that the accuracy stays steady at high value for almost all δ from 0 to 0.3 and that the survival model with smaller smoothing window exhibits higher prediction accuracy. In addition, the average transitional period duration Δt against different smoothing length is calculated as well. Each values of τ corresponds to a curve in Figure 3 that stands for the average transitional period duration; survival model with lower value of τ has shorter transitional period, which means better performance.

It, however, seems contradictory that narrower smoothing window leads to better performance while the calculated appropriate length existed among a moderate value. Considering the defects a small τ brings as mentioned in Section II, consequently, we compute the amount of reoccurrence friendships among various window length. We summarize the results in Figure 4, from which we could clearly see that the friendships seem to be less probable to resurrect after having be terminate when we adopt a wider smoothing window. In summary, there does exist a tradeoff between the accuracy of predicting surviving time for each online friendships and the total times of online friendship reoccurring in whole social network, which makes medium window length more appropriate.

V. Conclusion And Future Work

In this paper, we explored the evolution of friendships in online social networks. We constructed a multivariate survival model with high accuracy of predicting surviving time of each friendships. By analyzing moving average process, we proposed a scientific method in choosing smoothing length which handles tradeoff between high accuracy and little friendship reoccurrence. Through evaluation on real data set, we presented details of our model performance.

For our future work, we will adopt more complicated kernel function such as anisotropic Gaussian radial basis function kernel. In addition, since the friendships between users are predicted precisely, in future, we will explore the information spreading pattern in online social network.

Acknowledgment

This work is supported by NSF China (No.61325012, 61271219, 61221001, 61202373, 61102052); SEU SKL project (No.2012D13); Shanghai Basic Research Key Project (No. 13510711300, 12JC1405200, 11JC1405100).

References

[1] L. Fu, J. Zhang, and X. Wang, "Evolution-cast: Temporal evolution in wireless social networks and its impact on capacity," 2013.

[2] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler, "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, no. 7415, pp. 295–298, 2012.

[3] S. G. Roberts, R. I. Dunbar, T. V. Pollet, and T. Kuppens, "Exploring variation in active network size: Constraints and ego characteristics," *Social Networks*, vol. 31, no. 2, pp. 138–146, 2009.

[4] V. Arnaboldi, M. Conti, A. Passarella, and F. Pezzoni, "Ego networks in twitter: an experimental analysis," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 3459–3464.

[5] G. Kossinets and D. J. Watts, "Empirical analysis of an evolving social network," *Science*, vol. 311, no. 5757, pp. 88–90, 2006.

[6] J. G. Lee, S. Moon, and K. Salamatian, "An approach to model and predict the popularity of online contents with explanatory factors," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1. IEEE, 2010, pp. 623–630.

[7] J. D. Hamilton, *Time series analysis*. Princeton university press Princeton, 1994, vol. 2.

[8] J. Chow, "On the estimation of the order of a moving-average process," *Automatic Control, IEEE Transactions on*, vol. 17, no. 3, pp. 386–387, 1972.

[9] D. R. Cox and D. Oakes, *Analysis of survival data*. CRC Press, 1984, vol. 21.

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[11] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Computational learning theory*. Springer, 2001, pp. 416–426.

[12] S. Fine and K. Scheinberg, "Efficient svm training using low-rank kernel representations," *The Journal of Machine Learning Research*, vol. 2, pp. 243–264, 2002.

[13] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.