Uncovering the Co-driven Mechanism of Social and Content Links in User Churn Phenomena

Yunfei Lu
Tsinghua University
luyf16@mails.tsinghua.edu.cn

Linyun Yu
Bytedance AI Lab
yulinyun@bytedance.com

Peng Cui
Tsinghua University
cuip@tsinghua.edu.cn

Chengxi Zang
Tsinghua University
zangcx13@mails.tsinghua.edu.cn

Renzhe Xu
Tsinghua University
xrz15@mails.tsinghua.edu.cn

Yihao Liu
Bytedance AI Lab
liuyihao@bytedance.com

Lei Li
Bytedance AI Lab
lilei.02@bytedance.com

Wenwu Zhu
Tsinghua University
wwzhu@tsinghua.edu.cn

ABSTRACT

Recent years witness the merge of social networks and user-generated content (UGC) platforms. In these new platforms, users establish links to others not only driven by their social relationships in the physical world but also driven by the contents published by others. During this merging process, social networks gradually integrate both social and content links and become unprecedentedly complicated, with the motivation to exploit both the advantages of social viscosity and content attractiveness to reach the best customer retention situation. However, due to the lack of fine-grained data recording such merging phenomena, the co-driven mechanism of social and content links in churn phenomena remains unexplored. How do social and content factors jointly influence customers' churn? What is the best ratio of social and content links for a user's retention? Is there a model to capture this co-driven mechanism in users' churn phenomena?

In this paper, we collect a real-world dataset with more than 5.77 million users and 925 million links, with each link being tagged as a social one or a content one. We find that both social and content links have a significant impact on users' churn and they work jointly as a complicated mixture effect. As a result, we propose a novel survival model, which incorporates both social and content factors, to predict churn probability over time. Our model successfully fits the churn distribution in reality and accurately predicts the churn rate of different subpopulations in the future. By analyzing the modeling parameters, we try to strike a balance between social-driven and content-driven links in a user's social network to reach the lowest churn rate. Our model and findings may have potential implications for the design of future social media.

KEYWORDS

Co-driven Mechanism of Social and Content Links; User Churn; Survival Analysis

1 INTRODUCTION

Social networks and user-generated content(UGC) platforms are two prominent types of online services, satisfying social demands and information demands respectively. The early formation of social media is an instant messaging tool with few other functions such as MSN and QQ, while inchoate UGC platforms only provide content and pay little attention to social relationships of users, such as Yahoo Answers. Some researches demonstrate that they can complement each other [24]. However, the merge of UGC and social networking is an important trend in recent years. Social service providers and content service providers take different paths. Mainstream social media like Facebook, Twitter, and WeChat, encourage people to post and share information over established social networks, which makes contributing and consuming content to be important functions of the social network. Rising popular UGC platforms, like Instagram and TikTok, not only help people establish connections with strangers through content-based community or recommendation systems, but also stimulate them to bring in their social relationships to the platform. The merge of social networks and UGC platforms produces unprecedentedly complicated network structure where social-driven links (i.e. the following relationships due to friendship) and content-driven links (i.e. the following relationships due to interests in generated content) coexist. The key motivation of the merge is to exploit the advantages of social viscosity and content attractiveness to reach the best customers' retention situation.

In literature, since the retention of a user is the basis of other behavior, such as information diffusion and group evolution, user churn is regarded as one of the most important issues in social networks [14]. Much effort has been made to model and predict churn, ranging from the feature-driven approaches with various user characteristics collected [7], to the influence-based techniques with the adoption of information diffusion models [15]. Besides predicting the result of churn only, some studies also try to model the probability of churn over time by combining the feature-driven method with the survival model [12]. The importance of relationships in

Conference'17, July 2017, Washington, DC, USA
© 2019 Association for Computing Machinery.
ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

a social network is also proved by previous studies [5]. However, none of them systematically analyze the co-driven mechanism of social and content-driven relationships in churn phenomena.

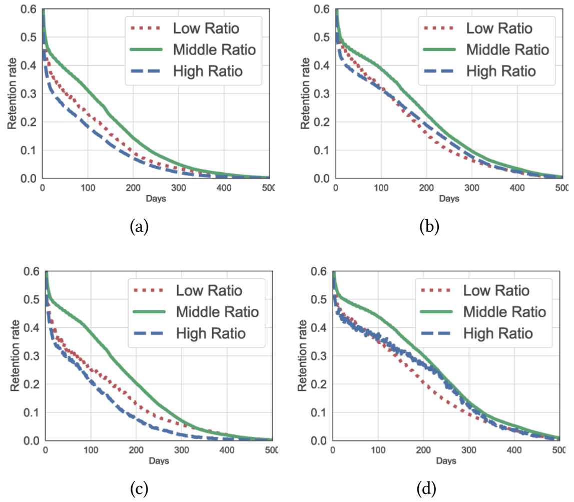
In this paper, we collect a large-scale and fine-grained dataset with more than 5.77 million users and 1.15 billion links spanning nearly two years from TikTok(Douyin), one of the most popular UGC platforms in China. The dataset explicitly records when and how a link is established. Specifically, one link is tagged as a social link if it is created through the mobile phone address book. If a link cannot be identified as a social link, and the involved two users get linked through content-based recommendation function provided by the platform, then the link is tagged as a content link. To the best of our knowledge, this is the first large-scale dataset that explicitly distinguishes content-driven links from social-driven links, enabling the research on the co-driven mechanism of these two factors in a social network.

By analyzing the effects of different types of links in the process of churn, we find rich patterns indicating that both social and content are important for users' retention. We compute social-content ratio for each user by dividing the number of social links by the sum of social and content links, in both inward and outward links for each user respectively. We plot the retention rate over time for users with different levels of social-content ratio in two user sets of different levels of activeness. As shown in Fig.1, users with a moderate social-content ratio in both inward and outward links achieve the highest retention rate on both user sets , which demonstrates that both social and content factors have a significant impact on churn and a balance between them leads to the lowest churn probability. It also indicates that social and content links influence churn as a complicated mixture effect rather than independently. Otherwise, if they are independent and have the same influence, the three curves in Fig. 1 should overlap; if assuming a simple linear mixture between social and content links, the lowest retention rate should be realized by either the highest or lowest social-content ratio. Both situations are contradictory with reality.

Based fundamentally on our finds, we propose a survival model, named Social-Content Mixture model(SCM), incorporating the effects of both social-driven and content-driven links, to capture the dynamic churn rate over time. The effectiveness of our SCM model is validated by fitting the distribution of churn rate and churn time of our large-scale empirical dataset. Furthermore, our model accurately predicts the churn event in the future. Last but not least, by analyzing the modeling parameters, we find the very ratio of social-driven and content-driven links which leads to the lowest churn rate. Our findings and our SCM model have potential implications for the design of the next social network and UGC platforms.

In summary, it is worthwhile to highlight our contributions as follows:

- **Novel Findings:** We find both social-driven and content-driven links have a significant impact on users' churn phenomena, indicating the necessity of modeling this co-driven mechanism.
- **Our Novel SCM model:** We propose a novel survival model which incorporates and disentangles this co-driven mechanism to model the dynamic churn rate over time. The parameters of our SCM model have clear physical meanings.



(a)

(b)

(c)

(d)

Figure 1: ***We discover that moderate social-content ratio leads to highest retention in real-world dataset. (a)(b) is a set of inactive users whose number of links are in [10, 50], while (c)(d) is a set of active users whose number of links are in [100, 500]. The curves of different colors represent the retention rate of users with different levels of social-content ratio in the number of outward links((a)(c)) and inward links((b)(d)). The users with moderate ratio achieve the highest retention in all the illustrations, which means social and content are both important in churn and they influence churn as a complicated mixture effect.***

- **Accuracy and usefulness:** Our SCM model fits the empirical churn rate accurately and predicts the churn in the future very well. By applying our model, we find the ratio of social-driven and content-driven links which leads to the lowest user churn rate.

The rest of the paper is organized as follows: Section 2 gives a survey on the related work. We propose our SCM model in Section 3. In Section 4, we evaluate our method and report the result. Last, we conclude our work in Section 5.

2 RELATED WORK

As the investigated problem is closely related to churn prediction and the main advantage of our work is studying the co-driven mechanism of heterogeneous relationship, we mainly review the related works in these two fields.

Churn prediction. Emerging from business spaces and broadcast providers as a major issue[11], churn is also regarded as a crucial problem in online communities[25], social networks[27] and discussion sites[22]. Traditional churn prediction has been approached as a feature-based problem, which usually consists of feature selection and prediction model[13]. Among them, Dasgupta et. al[5] indicate that social relations play an important role in affecting customer churn. Althoff et.al[1] find that positive interactions can reduce churn significantly. Dave et.al[6] propose time-spent based models for user retention prediction. Other methods believe that some key individuals may influence others to churn themselves through network structure, for which they regard churn as a kind of cascade and adopt diffusion models for prediction[14]. Recently, some studies go further to predict the probability of churn with time. Zhu et. al[27] develop a personalized and socially regularized time-decay model to predict user activeness. Kapoor et. al[12] combine features and survival analysis through Cox proportional hazard model to predict the return time of users.

However, none has uncovered the co-driven mechanism of social and content links in churn and provided corresponding models due to lack of data that distinguishes these two different relations.

Heterogeneous relationships analysis. There are mainly two kinds of studies that divide relationships into different types and explore the joint influence from them on various problems of social networks. The first kind distinguishes reciprocal links from others. Hidalgo et. al[8] find that reciprocity is the strongest predictor of link persistence. Hopcroft et .al[10] prove the existence of the structural balance among reciprocal relationships. Cai et .al[2] provide a novel framework to learn reciprocal knowledge for link prediction. The second kind divide links into positive and negative ones depending on whether it expresses a positive or negative attitude[18]. Kunegis et al.[17] show that the allowance of establishing negative link has a small but measurable added value for social networks. Song et .al[23] propose a generalized metric to quantify the ranking performance of recommendation in networks with both positive and negative links. There are also studies on bringing in external links and explore their influence[16].

However, our novel findings in the network integrating social and content links are different from all the heterogeneous relationships analysis above and there are no survival or dynamic models that can capture the co-driven mechanism of these heterogeneous relationships.

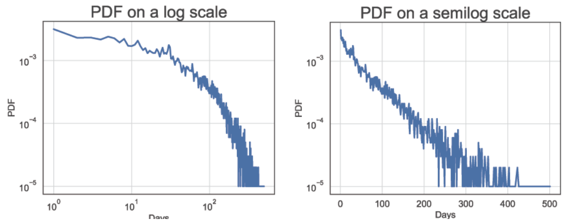
3 PROPOSED METHOD

In this section, we present our proposed model and analyze it. In order to depict the churn probability over time, we set up our model on the basis of survival analysis[4] and point process. In survival analysis, $f(t)$ denotes the probability density function which records the probability that an event happens at t. $S(t)$ denotes the probability that the event did not happen before t. $\lambda(t)$ denotes the conditional probability that the event will occur at t if it did not happen before t, which is called hazard function or intensity function. Given one of the three functions above, the other two can be determined by the following equations:

$$
\begin{aligned}
S(t) &= \int_t^{\infty} f(\tau)d\tau = e^{-\int_0^t \lambda(\tau)d\tau} \\
f(t) &= -\frac{\partial S(t)}{\partial t} = \lambda(t)e^{-\int_0^t \lambda(\tau)d\tau} \\
\lambda(t) &= \frac{f(t)}{S(t)} = \frac{f(t)}{\int_t^{\infty} f(\tau)d\tau}
\end{aligned}
\tag{1}
$$

We conduct survival analysis on the churn events of users, and choose to model the hazard function since it usually has the most succinct formulation. Next, we will describe our model in steps, adding complexity, and start with a null model.

3.1 Null model



(a) (b)

Figure 2: ***The probability distribution function(PDF) of churn time in real-world data on a log-log scale(a) and a semilog scale(b), which approximately complies with stretched exponential distribution with a cutoff.***

Since the hazard function describes the intensity of churn over time, we begin with the analysis of the probability distribution function(PDF) of churn time in real-world data. As shown in Fig.2(a), the PDF curve is far from straight at a log-log scale, which means it is not power law distribution. However, in Fig.2(b), the PDF curve is relatively straight but has a cutoff at the tail. Although a straight line on a semilog scale means exponential distribution, exponential distribution can not generate a cutoff since it is time-independent. Inspired by the work of Zang et al.[26], we adopt the stretched exponential distribution with a cut off as our null model:

$$
\lambda(t) = \frac{c}{t^{\theta}} + b \tag{2}
$$

where $\theta \neq 1$. Stretched exponential distribution is also known as Weibull distribution. When $\theta = 0$, $\lambda(t)$ will become a constant and generate an exponential distribution.

3.2 The SCM model

Here we propose the Social-Content Mixture(SCM) model based on the formation and analysis from our null model. Since the links in our dataset are directed, the links of each user can be divided into outward links and inward links. The outward links generate when one user follows others, whereas the inward links emerge when others follow him/her. Accordingly, the out-degree refers to the number of the outward links from one user, and the in-degree denotes to the number of the followers (i.e. fans). With social and content distinguished, we can divide the links into four types: social outward links, content outward links, social inward links, and content inward links, as shown in Fig.3. These four types of links for one user change over time before churn, for which we use a tetrad $X_u(t) = \{x_{u,s,in}(t), x_{u,c,in}(t), x_{u,s,out}(t), x_{u,s,out}(t)\}$ to represent the number of each type of link for u at t.

As shown in equation 2, b represents the factors that are time-independent while c incorporates the time-related factors of churn. The number of social and content links are keep changing over the life cycle of each user, for which the influence from links should be incorporated in c. Since we focus on exploring the effect of the four types of links, we decompose c into two parts: the function of

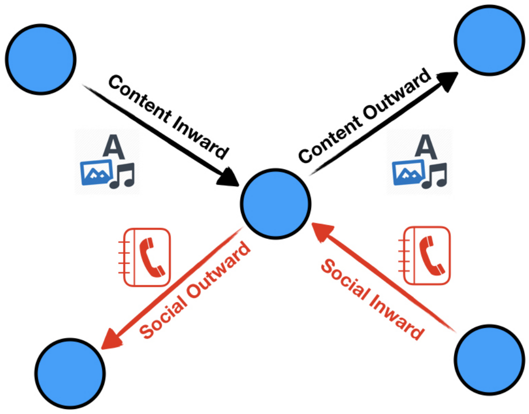


Figure 3: ***An illustration of the four different types of links in the network of TikTok(Douyin).***

$X_u(t)$ and the other time-varying factors a that are independent of $X_u(t)$. So we have

$$c = ag(X_u(t)) = ag(x_{u,s,out}(t), x_{u,c,out}(t), x_{u,s,in}(t), x_{u,c,in}(t)) \tag{3}$$

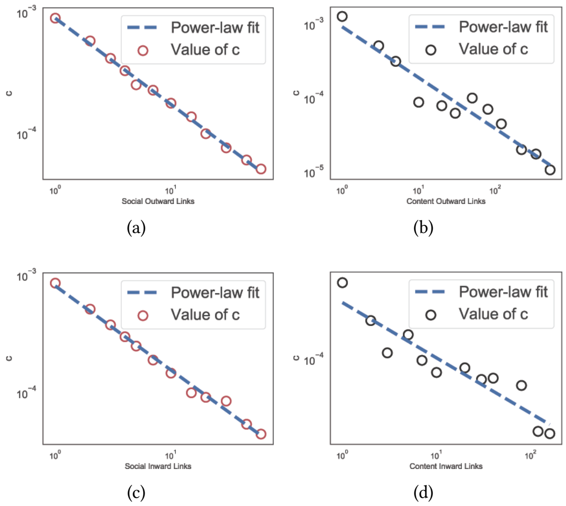
The analysis from former researches on directed links demonstrates that following and being followed have different forms of effects to users[9]. Following decides what will see, which satisfies the demands of consuming contents, while being followed satisfies the demands of sharing and achieving a sense of achievement. Since the mechanisms of following and being followed are relatively unrelated, we regard the effect of inward and outward as mutually independent. So we can decompose equation 3 as follow:

$$c = ag(X_ut) = sg_1(x_{u,s,out}(t), x_{u,c,out}(t))g_2(x_{u,s,in}(t), x_{u,c,in}(t)) \tag{4}$$

As for the relation of social and content links with the same direction, from Fig. 1 we can see, the two kinds of links influence churn as a mixture effect rather than independently. The most intuitive and succinct formation of a mixture model is to add each component together, which changes equation 4 to

$$c = a(g_1(x_{u,s,out}(t)) + g_2(x_{u,c,out}(t)))(g_3(x_{u,s,in}(t)) + g_4(x_{u,c,in}(t))) \tag{5}$$

The last step is to figure out the function form of the number of each type of links. Since more links usually indicate a lower probability of churn, it is likely that c decreases with the number of links, for which $g(x)$ decreases with x. We first focus on $x_{u,s,out}$ and collect a set of users $S_{s,out}$, where they have different numbers of $x_{u,s,out}$ and same numbers of other three types of links. For the user in $S_{s,out}$, we can rewrite equation 6 to $c = \gamma(g_1(x_{u,s,out}) + \beta)$, where $\gamma = a(g_3(x_{u,s,in}(t)) + g_4(x_{u,c,in}(t)))$ and $\beta = g_2(x_{u,c,out})$ are constant. In order to highlight the effect of $x_{u,s,out}$, we hope β to be as small as possible, for which we only choose the users with big $x_{u,c,out}$ to lower $\beta = g_2(x_{u,c,out})$. After that, we separate users with $x_{u,s,out} = i$ in $S_{s,out}$ into $S_{s,out,i}$ and train our null model on each $S_{s,out,i}$. We learn the parameters $c_{s,out,i}$ from $S_{s,out,i}$ and plot the relation between $c_{s,out,i}$ and i. We repeat the process above for the other three types of links and display the curves in Fig. 4.



(a)

(b)

(c)

(d)

Figure 4: ***The relation between*** c ***in null model and the number of (a)social outward links (b)content outward links (c)social inward links (d) content outward links on a log-log scale. The numbers of four types of links seem to have power law influence on churn.***

From Fig. 4 we can see, the relation between c and four types of links looks like a straight line on a log-log scale, which means $\log c \approx k \log x + m$ and indicates power law influence. Combined with $c = \gamma(g(x) + \beta)$ we get

$$\begin{aligned} \log(\gamma g(x) + \gamma\beta) &\approx k \log x + m \\ g(x) &\approx e^m x^k - \gamma\beta \end{aligned} \tag{6}$$

where m is the intercept in Fig. 4. Since the value of $\gamma\beta$ and m are both very close to 0, we have $g(x) \approx x^k$. Substitute the above results in equation 4 then we have

$$c = a(x_{u,s,out}^{k_{s,out}}(t) + x_{u,c,out}^{k_{c,out}}(t))(x_{u,s,in}^{k_{s,in}}(t) + x_{u,c,in}^{k_{c,in}}(t)) \tag{7}$$

Combining equation 7 with our null model, we finally get the form of our Social-Content Mixture(SCM) model.

$$\lambda(t) = \frac{a(x_{u,s,out}^{k_{s,out}}(t) + x_{u,c,out}^{k_{c,out}}(t))(x_{u,s,in}^{k_{s,in}}(t) + x_{u,c,in}^{k_{c,in}}(t))}{t^{\theta}} + b \tag{8}$$

From Fig. 4 we can see that $k < 0$ for all the four types of links.

Justification of the model:

Influence from links. $k_{s,out}$, $k_{c,out}$, $k_{s,in}$ and $k_{c,in}$ capture the influence from different types of links. Social and content links with the same direction work as a mixture, leading to the phenomenon that moderate social-content ratio achieves the lowest churn rate in Fig. 1. For example, if u has many content outward links but few social outward links at t, since $k < 0$, $x_{u,c,out}^{k_{c,out}}(t) \rightarrow 0$ with the big number of $x_{u,c,out}(t)$, the factors from outward links

$(x_{u,s,out}^{k_{s,out}}(t) + x_{u,c,out}^{k_{c,out}}(t))$ will be dominated by $x_{u,s,out}^{k_{s,out}}(t)$, for which further growth of content outward links is useless but more social outward links will have an evident effect on reducing the probability of churn. In a similar way, if u has much more social links than content links with the same direction, the number of content links will become the key factor of churn. Thus, a balance between social and content links leads to the highest retention rate, which is identical to our observation in real-world dataset.

Influence from other time-varying factors. a describes the average effect of time-varying factors except the number of links, such as the interest of users to the platform. A small value of a will lead to low churn probability even if there are few links. When a is big, the intensity of churn depends on the influence from links.

Time effect. θ indicates the impact of time itself over churn, excluding the influence from time-varying factors. $\theta > 0$ means the probability of churn decays with time increasing, indicating the long-term dependence and stickiness of users on the platform. When $\theta = 0$, time itself is independent with churn and the null model will degenerate to Poisson process. If $\theta < 0$, the probability of churn increases with time and those users without growth on their links will gradually leave the platform.

Time-independent effect. b captures the effect from time-independent factors of churn, such as sex ratio, age distribution, quality of recommendation system and so on. At the initial stage of use, users do not have many links and the time decay effect is still weak, for which b plays the most important role at short-term scale. If b is high, lots of users will churn early, which is common among unsuccessful platforms. If b is low, more users can survive the initial stage and may obtain enough links to achieve high retention.

3.3 Parameter estimation

The parameters of our null model and SCM model can be learned by the maximum likelihood estimation(MLE) framework. Our parameter set consists of $\{c, \theta, b\}$ for null model and $\{a, \theta, b, k_{s,out}, k_{c,out}, k_{s,in}, k_{c,in}\}$ for SCM model.

3.3.1 Estimation of null model. Let t_u be the last time that u appears in the dataset. For the users who churn in the observation period, we know the churn time for each user and denote the set of churn users as CS(churn set). The likelihood for u in CS who churn at t_u is $f(t_u)$. Other users, denoted by RS(retention set), have not churned at the end of the observation or become invisible halfway. We only know that the churn time of u is bigger than t_u, whose likelihood is $S(t_u)$. So the likelihood function of all the users is:

$$\begin{aligned} L &= \prod_{u \in CS} f(t_u) \prod_{u \in RS} S(t_u) = \prod_{u \in CS} \lambda(t_u) \prod_{u} S(t_u) \\ &= \prod_{u \in CS} \left(\frac{c}{t_u^{\theta}} + b \right) \prod_{u} e^{-\frac{ct_u^{1-\theta}}{1-\theta} - bt_u} \\ \log L &= \sum_{u \in CS} \log \left(\frac{c}{t_u^{\theta}} + b \right) + \sum_{u} \left(-\frac{ct_u^{1-\theta}}{1-\theta} - bt_u \right) \end{aligned} \tag{9}$$

We calculate the gradients for all the parameters as follow:

$$\begin{aligned} \frac{\partial \log L}{\partial c} &= \sum_{u \in CS} \frac{1}{c + bt_u^{\theta}} + \sum_{u} (-\frac{t_u^{1-\theta}}{1-\theta}) \\ \frac{\partial \log L}{\partial \theta} &= \sum_{u \in CS} \left(-\frac{c \log t_u}{c + bt_u^{\theta}} \right) + \sum_{u} \left(-\frac{ct_u^{1-\theta}(1 - (1-\theta) \log t_u)}{(1-\theta)^2} \right) \\ \frac{\partial \log L}{\partial b} &= \sum_{u \in CS} \frac{1}{ct_u^{-\theta} + b} + \sum_{u} (-t_u) \end{aligned} \tag{10}$$

Then we learn the parameters with gradient descent optimizers.

3.3.2 Estimation of SCM model. SCM model incorporates the numbers of links of different types for each user, which change over time, making it very complicated to calculate $S(t) = e^{-\int_0^t \lambda(\tau) d\tau}$. There are mainly two methods to handle these time-varying covariates in survival analysis. The first one is to separate the records of each user into multiple intervals, making all the covariates remain constant during each interval, and calculate the likelihood based on these intervals[21]. The second one is to replace the original likelihood with partial likelihood[3], which is widely used in Cox proportional hazard model[20] since the parameters unrelated to covariates can be divided out. However, because of the existence of θ and b in SCM model, the form of partial likelihood will only lead to more complicated form, for which we choose the first method.

Denoting the i_{th} interval of u by $\Delta_{u,i}$ and its length by $t_{u,i}$, if u churns at the end of this interval, we put $\Delta_{u,i}$ into churn set(CS). Otherwise, $\Delta_{u,i}$ is placed into retention set(RS). The likelihood function of all the intervals is as follow:

$$\begin{aligned} L &= \prod_{\Delta_{u,i} \in CS} f(t_{u,i}) \prod_{\Delta_{u,i} \in RS} S(t_{u,i}) = \prod_{\Delta_{u,i} \in CS} \lambda(t_{u,i}) \prod_{\Delta_{u,i}} S(t_{u,i}) \\ &= \prod_{\Delta_{u,i} \in CS} \left(\frac{aR}{t_{u,i}^{\theta}} + b \right) \prod_{\Delta_{u,i}} e^{-\frac{aRt_{u,i}^{1-\theta}}{1-\theta} - bt_{u,i}} \\ \log L &= \sum_{\Delta_{u,i} \in CS} \left(\log(\frac{aR}{t_{u,i}^{\theta}} + b) \right) + \sum_{\Delta_{u,i}} \left(-\frac{aRt_{u,i}^{1-\theta}}{1-\theta} - bt_{u,i} \right) \end{aligned} \tag{11}$$

where $R = (x_{u,i,s,out}^{k_{s,out}} + x_{u,i,c,out}^{k_{c,out}})(x_{u,i,s,in}^{k_{s,in}} + x_{u,i,c,in}^{k_{c,in}})$ is constant in each interval. So the gradients for a, θ and b are similar to the forms in null model and can be figured out by replacing c with aR in equation 10 . The forms of gradients for four different k are similar and symmetric, for which we only list the gradients for $k_{s,out}$ for brevity.

$$\begin{aligned} \frac{\partial \log L}{\partial k_{s,out}} = &\sum_{\Delta_{u,i} \in CS} \frac{ax_{u,i,s,out}^{k_{s,out}} \log x_{u,i,s,out} (x_{u,i,s,in}^{k_{s,in}} + x_{u,i,c,in}^{k_{c,in}})}{ax_{u,i,s,out}^{k_{s,out}} (x_{u,i,s,in}^{k_{s,in}} + x_{u,i,c,in}^{k_{c,in}}) + (x_{u,i,c,out}^{k_{c,out}} + b) t_{u,i}^{\theta}} \\ &+ \sum_{\Delta_{u,i}} \left(-\frac{at_{u,i}^{1-\theta} x_{u,i,s,out}^{k_{s,out}} \log x_{u,i,s,out} (x_{u,i,s,in}^{k_{s,in}} + x_{u,i,c,in}^{k_{c,in}})}{1-\theta} \right) \end{aligned} \tag{12}$$

For the purpose of finding a good region of parameter space and getting faster convergence, we set the initial value of parameters using the result from null model and some prior knowledge from empirical data.

4 EXPERIMENTS

In this section, we evaluate the effectiveness of SCM model on the real data. We introduce our dataset in Sec.4.1 and the baseline models used in experiments in Sec.4.2. Sec.4.3 shows the accuracy of our model on fitting the distribution of churn in reality. In Sec.4.4, we demonstrate the predicting power of our model on the churn rate in the future. We give an insight into the balance points between social and content links with the parameters learned by SCM in Sec.4.5.

4.1 Datasets

Our experiments are conducted on a large-scale fine-grained dataset from TikTok(Douyin)[1]. It is one of the most popular UGC platforms of short-form mobile videos in China and owns 500 million monthly active users. The dataset explicitly records the time of each login for every user, and when and how a link is established, for 584 days from Jan 1, 2017 to Aug 8, 2018. Each link is tagged as a social link or content link based on its creation method. TikTok(Douyin) has an excellent content delivery system, leading to a majority of content links. Meanwhile, it also encourages users to bring in their social relationships. We sample 5.77 million users with 1.15 billion links from the users whose number of inward links are between 50 and 500 at Aug 8 2018 in our dataset. To our best knowledge, this is the first large-scale dataset that distinguishes social and content links and enables the research on the co-driven mechanism of these two most important factors in social network.

We regard the point when a user has over 10 inward links or outward links as the start point for the user, and define churn for a user if he or she does not log in for over 30 consecutive days, which is consistent with the real setting in Douyin platform.

4.2 Baselines for experiments

To exemplify the performance of our SCM model, we use some survival based models as baselines. The details of these models are described as follows:

1) Cox proportional hazard model (Cox)[20]: Cox model is one of the most frequently used multiple factor analysis methods in survival analysis. The hazard function for Cox model has the form

$$\lambda(t) = \lambda_0(t)e^{\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n} = \lambda_0(t)exp(x \cdot \beta) \qquad (13)$$

where x_i is a covariate and β_i is the corresponding weight. Here we adopt the numbers of four different types of links as the covariates. The baseline hazard function $\lambda_0(t)$ is estimated with the Breslow's method[19].

2) Cox model with log covariates (Cox-log): Cox model has an exponential form over the covariates. However, our observation in Fig.4 shows that the relations between hazard rate and the four numbers of links should be closer to power law. So we take the logarithm of the numbers of links as the covariates for this baseline.

3) Our null model (Null): Compared with SCM model, our null model lacks the influence from different types of links but has all the other mechanism.

4) Null model with product of covariates (Prod): Based on the null model, assumes that different types of links influence churn independently, and then the four x^k should be multiplied together, forming this baseline.

5) Null model with sum of covariates (Sum): Except for the independent assumption, another intuitive method to combine the four covariates is to simply linearly combine all the x^k together, which we adopt as the fifth baseline.

4.3 Accuracy

We validate the accuracy of our SCM model by answering if it can capture the churn rate distribution over time and over the numbers of different types of links. We separate our dataset into two sets with an equal amount of users and the same length of observation, one as the train set and the other as the test set. We train SCM model and the baselines on the train set and compute the expectation of churn rate over time on the test set with parameters learned.

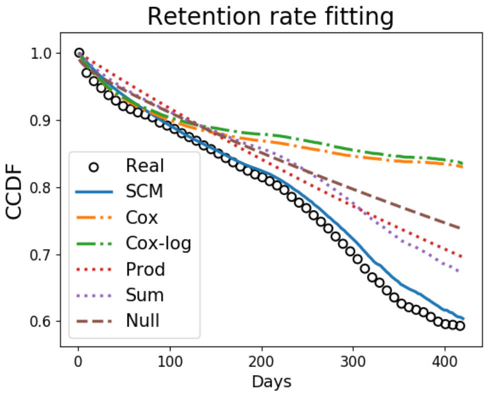


Figure 5: *Our model accurately fits the CCDF of churn, which means retention rate, superior to all the baseline models.*

Fig. 5 plots the complementary cumulative distribution function of churn in the test set, which is equivalent to retention rate and usually attracts major attention. Our model depicts the process of churn successfully and fits the retention rate accurately over the whole observation period, superior to all the baselines.

In order to verify the capacity of capturing the influence from different types of links, we divide the users into multiple subpopulations according to these numbers. Then we plot the heat map of churn rate on these subpopulations at $t = 100, 200, 300, 400, 500$. Fig. 6 is a showcase for the distribution over outward links and inward links respectively at $t = 100$. We can see that SCM model captures the co-driven mechanism of social and content links and fits the churn rate distribution well, except for some randomness in the real churn rates due to the limited population size. Since null model does not incorporate any covariates, we ignore it in this part.

We use Kolmogorov-Smirnov statistics(KS-stat) to quantitatively evaluate the performance of fitting CCDF. It is a frequently-used

[1] https://www.douyin.com/

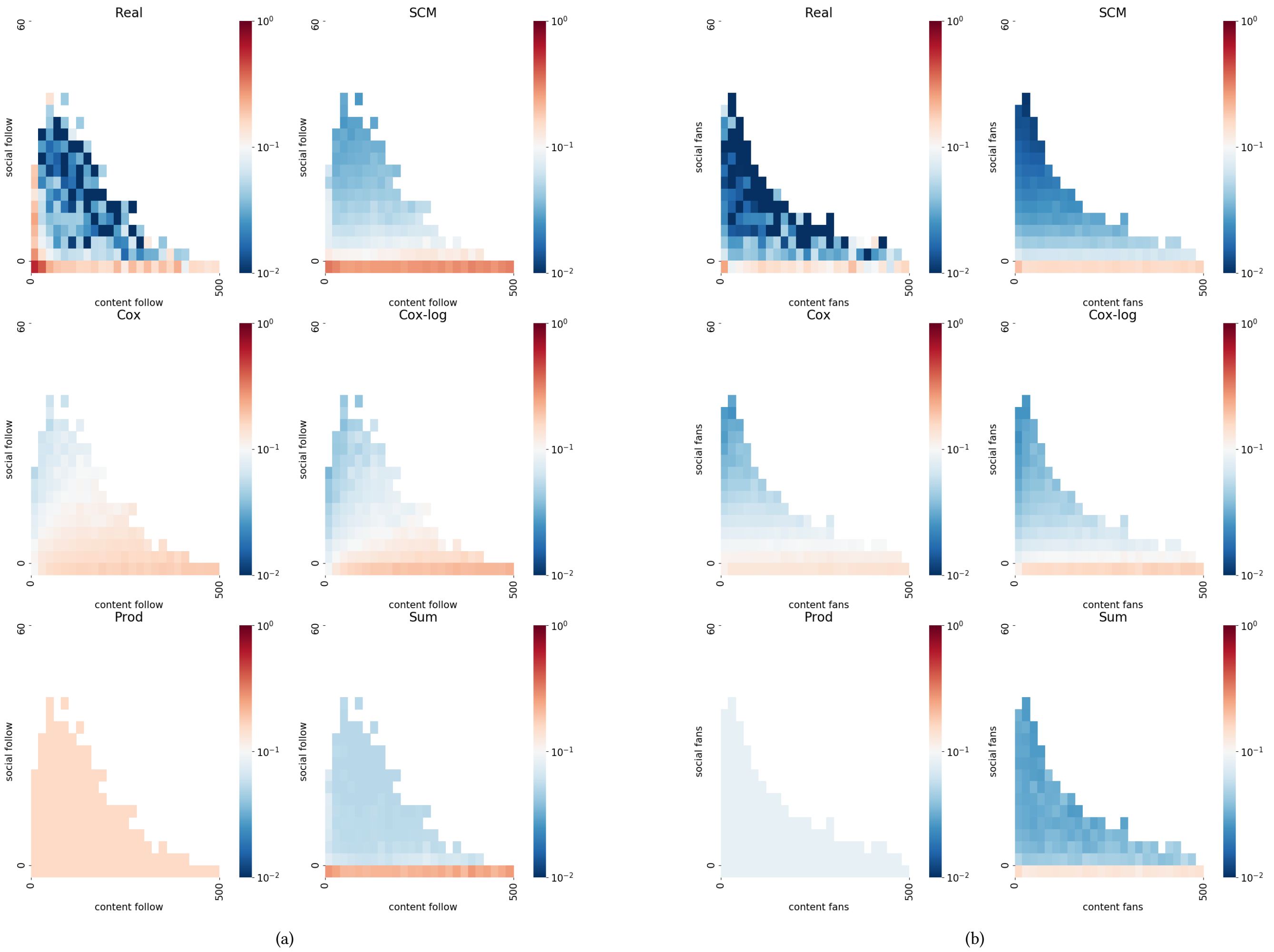


Figure 6: *The heat map of churn rate on social outward links vs. content outward links(a) and social inward links vs. content inward links(b) at* $t = 100$*. The color represents the churn rate of the corresponding subpopulation. Our model fits the churn rate distribution over different types of links pretty well except for some randomness and is the most closer one to reality.*

methods for comparing two subsets by quantifying a distance between the cumulative distribution functions of two samples. If the KS-stat is small or the p-value is high, the hypothesis that the distributions of the two subsets are the same cannot be rejected. KS-stat is calculated by

$$ksstat = sup_x |F_1(x) - F_2(x)| \tag{14}$$

As for the churn rate distribution over the four covariates, we use weighted mean absolute error(WMAE) as follow:

$$WMAE(t) = \frac{\sum_p w_p(t)|R_p(t) - \widehat{R_p(t)}|}{\sum_p w_p(t)} \tag{15}$$

where $w_p(t)$ is the size of subpopulation p at t. $R_p(t)$ is the churn number of p at t in real data, while $\widehat{R_p(t)}$ is the result from models. We denote WMAE on the distribution of outward links and inward links by WMAE-O and WMAE-I respectively, and list the fitting results from all the models in Table 1.

Table 1: Fitting results on the test set. Winner in bold.

Metrics	SCM	Cox	Cox-log	Null	Prod	Sum
KS-stat	**0.088**	0.579	0.593	0.348	0.279	0.245
WMAE-O(100)	**14.460**	21.781	21.193	-	25.177	14.673
WMAE-O(200)	**14.236**	24.194	23.048	-	24.867	14.952
WMAE-O(300)	**14.298**	31.921	31.337	-	27.943	16.963
WMAE-O(400)	**11.081**	37.892	37.774	-	28.441	16.434
WMAE-O(500)	**2.550**	19.927	19.942	-	13.278	5.029
WMAE-I(100)	**12.967**	32.193	28.955	-	37.430	17.207
WMAE-I(200)	**18.750**	32.254	30.663	-	34.669	21.250
WMAE-I(300)	**16.496**	33.870	32.861	-	32.403	19.928
WMAE-I(400)	**12.164**	31.024	30.796	-	26.190	15.903
WMAE-I(500)	**12.320**	65.097	64.986	-	45.753	21.315

In KS-stat, only our SCM model achieves a p-value of 0.073 that pass the test at the default 5% significance level, while the p-values of other metrics are all smaller than 10^{-8}. It means that all the baselines fail to capture the hazard rate of churn but SCM is consistent with the reality.

As for WMAE, the majority of worst results are reported by Cox and Cox-log. Although they incorporate covariates, they fail to estimate the hazard function due to their inadequate function form in depicting the process of churn. Prod also reports a large error, indicating the effects of social and content links are not completely independent. Although performing much better than Prod, Sum cannot generate the extreme low or high values in reality, such as the top part of the heat map, which is because that Sum simply combines all the effects linearly and tend to generate mean field results. However, our model incorporates the co-driven mechanism from all the different types of links successfully and achieves the best result in all the metrics.

4.4 Prediction

As a dynamic model, SCM model can be used to predict the churn rate in the future, which is a hot topic potentially leading to many valuable applications. In this section, we use the data of the first 145 days for training, with the purpose of predicting the retention rate and churn rate distribution in the future 438 days.

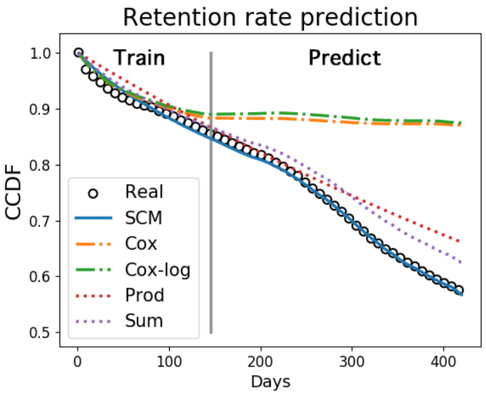


Figure 7: *Our model predicts the retention rate in the future accurately, significantly better to all the baseline models. For better visualization effect, we omit the result from null model which diverges from the reality a lot.*

As shown in Fig. 7, our SCM model outperforms all the baselines significantly. With only the first 25% of the period for training, SCM model precisely forecasts the churn events in the future. We also calculate the KS-stat and WMAE on the prediction results from the models, which is listed in Table 2. On KS-stat, our model achieves a p-value reaching up to 0.875, far exceeding the significance level of 5%, which means the results of SCM are very close to reality. Again, the p-values of other metrics are smaller than 10^{-8}. As for WMAE, SCM model beats all the baselines and the superiority over other models increases with the length of prediction.

Table 2: Fitting results on the test set. Winner in bold.

Metrics	SCM	Cox	Cox-log	Null	Prod	Sum
KS-stat	**0.040**	0.702	0.714	0.843	0.221	0.145
WMAE-O(200)	14.020	22.944	22.171	-	24.064	**13.827**
WMAE-O(300)	**13.519**	32.519	32.287	-	27.760	14.736
WMAE-O(400)	**9.853**	40.301	40.451	-	27.161	12.995
WMAE-O(500)	**3.235**	20.407	20.393	-	13.055	4.266
WMAE-I(200)	**20.677**	31.911	30.612	-	37.834	22.185
WMAE-I(300)	**17.985**	35.648	34.663	-	34.878	19.618
WMAE-I(400)	**12.193**	31.805	31.639	-	26.545	14.080
WMAE-I(500)	**17.919**	115.190	115.325	-	76.386	26.435

4.5 Insight

The usefulness of SCM model is not only embodied in its prediction power, but also reflected in the insight about link type suggestion. Here we consider a practical problem: given the number of social links and content links of a user, what kind of link should we recommend to him/her to decrease the churn probability? Which kind of people should we recommend him/her to?

Assume we are recommending user u to follow someone, namely, to establish a new outward link. Based on SCM model, the partial derivatives of λ with respect to $x_{u,s,out}$ and $x_{u,c,out}$ are

$$\begin{aligned}\frac{\partial \lambda}{\partial x_{u,s,out}} &= \frac{aR_{u,in}k_{u,s,out}x_{u,s,out}^{k_{u,s,out}-1}}{t^{\theta}} \\ \frac{\partial \lambda}{\partial x_{u,c,out}} &= \frac{aR_{in}k_{u,c,out}x_{u,c,out}^{k_{u,c,out}-1}}{t^{\theta}}\end{aligned} \tag{16}$$

where $R_{u,in} = x_{u,s,in}^{k_{u,s,in}} + x_{u,c,in}^{k_{u,c,in}}$. Since $k < 0$, both the partial derivatives are negative, which means any increase on these links will reduce the churn probability. However, the amounts of decreasing are different. As shown in 16, if $|k_{u,s,out}x_{u,s,out}^{k_{u,s,out}-1}|$ is bigger than $|k_{u,c,out}x_{u,c,out}^{k_{u,c,out}-1}|$, one social link will be more effective; otherwise, a content link will be the better choice. Once the parameters k are learned, it will be feasible to judge which kind of link to recommend for each user according to his/her current status. In a similar way, we can figure out which kind of inward links u needs more and make decisions of recommending u to whom.

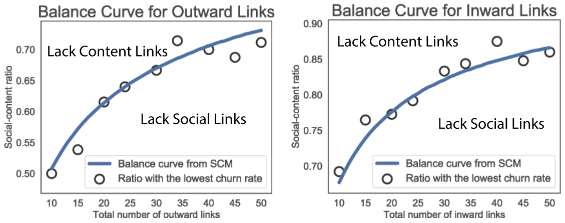
Note that $k < 0$, for which $|kx^{k-1}|$ will be increasingly smaller with the growth of x. If one already has lots of links of a certain type, further links of this type will be meaningless for retention. So given the total number of links, what is the best social-content ratio leading to the lowest churn probability? How to depict the balance points between social and content?

Assuming the total number of outward links is n and the social-content ratio in n is r, we calculate the partial derivative of λ with respect to r as follow:

$$\begin{aligned}\frac{\partial \lambda}{\partial r} &= \frac{anR_{u,in}}{t^{\theta}}(k_{u,s,out}(rn)^{k_{u,s,out}-1} - k_{u,c,out}((1-r)n)^{k_{u,c,out}-1}) \\ &= \frac{anR_{u,in}}{t^{\theta}}h(r)\end{aligned} \tag{17}$$

$h(r)$ is monotone increasing with r on (0, 1). Since $h(r) < 0$ for $r \rightarrow 0$ and $h(r) > 0$ for $r \rightarrow 1$, $h(r)$ has a root in (0, 1) where λ

reaches its minimal point and the root is the balance point leading to the lowest churn probability. We substitute the value of k learned from our dataset into $h(r)$, and calculate the balance ratio for each n up to 50, since the users with links less than 50 are most likely to churn in our dataset. We make up a curve with those points and compare it with the social-content ratio that has the lowest churn rate in real-world data.



(a)

(b)

Figure 8: ***The balance curves calculated from our model are very close to the social-content ratio with the lowest churn probability in reality. (a) is for outward links and (b) is for inward links.***

As shown in Fig.8, the balance curves suggested by our model is very close to the social-content ratio with the lowest churn rate in real data for both inward and outward links, which again demonstrates the accuracy of SCM. The balance curves are of high significance which may serve as a guide for link recommendation. Approaching the balance curve leads to lower churn probability, for which the users above the curve need to establish more content links, while the users below the curve should be encouraged to establish social relationships on this platform.

5 CONCLUSIONS

In this paper, we study the co-driven mechanism of social and content links on a large-scale and fine-grained real-world dataset, where each link is tagged as a social one or a content one . We find that social links and content links influence user churn as a complicated mixture effect rather than independently. Based on our exploration, we propose a survival model named Social-Content Mixture model(SCM), to incorporate both social and content factors. Our model achieves high accuracy on fitting the churn distribution in reality and predicting the churn rate in the future. With the modeling parameters, we find the ratio of social and content links which leads to the lowest churn rate. The main contributions are:

- **Novel Findings:** We find both social and content links have an significant impact on user churn phenomena, indicating the necessity of modeling this co-driven mechanism.
- **Our Novel SCM model:** We propose a survival model to incorporate this co-driven mechanism and model the churn rate over time. SCM model has a succinct form and each parameter has clear physical meanings.
- **Accuracy and usefulness:** SCM model fits the churn distribution in reality pretty well and predicts the churn rate in the future accurately. By applying our model, we find the ratio of social-driven and content-driven links which leads to the lowest churn probability.

REFERENCES

[1] T Althoff and J Leskovec. 2015. Donor Retention in Online Crowdfunding Communities: A Case Study of DonorsChoose.org.
[2] Xiongcai Cai, Michael Bain, Alfred Krzywicki, Wayne Wobcke, Sok Kim Yang, Paul Compton, and Ashesh Mahidadia. 2012. Reciprocal and Heterogeneous Link Prediction in Social Networks. *Advances in Knowledge Discovery and Data Mining* 7302 (2012), 193–204.
[3] David R Cox. 1975. Partial likelihood. *Biometrika* 62, 2 (1975), 269–276.
[4] David Roxbee Cox. 2018. *Analysis of survival data.* Routledge.
[5] Koustuv Dasgupta, Rahul Singh, Balaji Viswanathan, Dipanjan Chakraborty, Sougata Mukherjea, Amit A. Nanavati, and Anupam Joshi. 2008. Social ties and their relevance to churn in mobile telecom networks. *Proc of Edbt* (2008), 668–677.
[6] Kushal Dave, Vishal Vaingankar, Sumanth Kolar, and Vasudeva Varma. 2013. Timespent Based Models for Predicting User Retention. In *International Conference on World Wide Web.*
[7] Gideon Dror, Pelleg Dan, Oleg Rokhlenko, and Idan Szpektor. 2012. Churn prediction in new users of Yahoo! Answers.
[8] Cesar A Hidalgo and Carlos Rodríguez-Sickert. 2008. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications* 387, 12 (2008), 3017–3024.
[9] Matthias Hofer and Viviane Aubert. 2013. Perceived bridging and bonding social capital on Twitter: Differentiating between followers and followees. *Computers in Human Behavior* 29, 6 (2013), 2134–2142.
[10] John E. Hopcroft, Tiancheng Lou, and Tang Jie. 2011. Who will follow you back? reciprocal relationship prediction. In *Acm International Conference on Information and Knowledge Management.*
[11] Shin-Yuan Hung, David C Yen, and Hsiu-Yu Wang. 2006. Applying data mining to telecom churn management. *Expert Systems with Applications* 31, 3 (2006), 515–524.
[12] Komal Kapoor, Mingxuan Sun, Jaideep Srivastava, and Ye Tao. 2014. A hazard based approach to user return time prediction. In *Acm Sigkdd International Conference on Knowledge Discovery and Data Mining.*
[13] Marcel Karnstedt, Jeff Chan, Conor Hayes, Harith Alani, and Matthew Rowe. 2011. The Effect of User Features on Churn in Social Networks. In *Web Science Conference.*
[14] Marcel Karnstedt, Tara Hennessy, Jeffrey Chan, and Conor Hayes. 2010. Churn in Social Networks: A Discussion Boards Case Study. In *IEEE Second International Conference on Social Computing.*
[15] Jaya Kawale, Aditya Pal, and Jaideep Srivastava. 2009. Churn prediction in MMORPGs: A social influence based approach. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, Vol. 4. IEEE, 423–428.
[16] Young Soo KIM, Kyungsub Stephen Choi, and Felicia Natali. 2016. Extending the network: The influence of offline friendship on Twitter network. AIS.
[17] JÃĺrÃťme Kunegis, Julia Preusse, and Felix Schwagereit. 2013. What is the added value of negative links in online social networks?
[18] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Predicting Positive and Negative Links in Online Social Networks. In *International Conference on World Wide Web.*
[19] DY Lin. 2007. On the Breslow estimator. *Lifetime data analysis* 13, 4 (2007), 471–480.
[20] Danyu Y Lin and Lee-Jen Wei. 1989. The robust inference for the Cox proportional hazards model. *Journal of the American statistical Association* 84, 408 (1989), 1074–1078.
[21] Teresa M Powell and Melissa E Bagnell. 2012. Your "survival" guide to using time-dependent covariates. In *Proceedings of the SAS Global Forum*, Vol. 2012. Citeseer, 22–25.
[22] Jagat Sastry Pudipeddi, Leman Akoglu, and Hanghang Tong. 2014. User churn in focused question answering sites: characterizations and prediction. *Sheridan-printing Com* (2014), 469–474.
[23] Dongjin Song and David A Meyer. 2015. Recommending Positive Links in Signed Social Networks by Optimizing a Generalized AUC.. In *AAAI.* 290–296.
[24] Qi Su and Wendell Baker. 2012. Access to trusted user-generated content using social networks. (Aug. 21 2012). US Patent 8,250,096.
[25] Jiang Yang, Xiao Wei, Mark S Ackerman, and Lada A Adamic. 2010. Activity Lifespan: An Analysis of User Survival Patterns in Online Knowledge Sharing Communities. *ICWSM* 10 (2010), 186–193.
[26] Chengxi Zang, Peng Cui, and Wenwu Zhu. 2018. Learning and Interpreting Complex Distributions in Empirical Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2682–2691.
[27] Yin Zhu, Erheng Zhong, Sinno Jialin Pan, Xiao Wang, Minzhe Zhou, and Qiang Yang. 2013. Predicting user activity level in social networks. In *Acm International Conference on Information and Knowledge Management.*