A Temporally Heterogeneous Survival Framework with Application to Social Behavior Dynamics

Linyun Yu
Department of Computer Science and Technology, Tsinghua University
East Main Building 9-316, Tsinghua University
Beijing, Beijing, China 100084
yuly12@mails.tsinghua.edu.cn

Peng Cui
Department of Computer Science and Technology, Tsinghua University
East Main Building 9-316, Tsinghua University
Beijing, Beijing, China 100084
cuip@tsinghua.edu.cn

Chaoming Song
Department of Physics, University of Miami
Rm 305, James L.Knight Physics Building, 1320 Campo Sano Ave
Coral Gables, Florida 33146, U.S.A
chaoming.song@gmail.com

Tianyang Zhang
Department of Computer Science and Technology, Tsinghua University
East Main Building 9-316, Tsinghua University
Beijing, Beijing, China 100084
zhangty09@foxmail.com

Shiqiang Yang
Department of Computer Science and Technology, Tsinghua University
FIT 3-512, Tsinghua University
Beijing, Beijing, China 100084
yangshq@tsinghua.edu.cn

ABSTRACT

Social behavior dynamics is one of the central building blocks in understanding and modeling complex social dynamic phenomena, such as information spreading, opinion formation, and social mobilization. While a wide range of models for social behavior dynamics have been proposed in recent years, the essential ingredients and the minimum model for social behavior dynamics is still largely unanswered. Here, we find that human interaction behavior dynamics exhibit rich complexities over the response time dimension and natural time dimension by exploring a large scale social communication dataset. To tackle this challenge, we develop a temporal Heterogeneous Survival framework where the regularities in response time dimension and natural time dimension can be organically integrated. We apply our model in two online social communication datasets. Our model can successfully regenerate the interaction patterns in the social communication datasets, and the results demonstrate that the proposed method can significantly outperform other state-of-the-art baselines. Meanwhile, the learnt parameters and discovered statistical regularities can lead to multiple potential applications.

KDD '17, August 13-17, 2017, Halifax, NS, Canada

DOI: http://dx.doi.org/10.1145/3097983.3098189

KEYWORDS

Social Dynamics, temporally Heterogeneous Survival framework, Human Behaviors

1 INTRODUCTION

Social behavior dynamics, referring to the dynamic process of human interactions, is one of the central building blocks in understanding and modeling complex social dynamic phenomena, such as information spreading, opinion formation, and social mobilization. While a wide range of models for social behavior dynamics have been proposed in recent years, most of them assume that the interactions among individuals are highly random, following a Poisson process [6, 20]. Recently, some recent non-trivial patterns of response time (i.e. the duration between the time a person receives a message and the time he makes a response) and the inter-event time (i.e. the time duration between consecutive behaviors of the same person) are found in empirical data [9, 23, 35]. A notable one [35] is that most responses made in a very short time scale, and some responses stall for a long time, resulting in a heavy-tailed distribution on response time, which is in contrast with the exponential distribution of response time generated by the Poisson process assumption on social interaction behaviors. Meanwhile, some recent study finds the evidence that the circadian rhythm [8], such as the working-rest periods, is non-trivial in influencing human behavior dynamics. What are the essential ingredients and the minimum model for social behavior dynamics is still largely unanswered.

Here, we explore a large scale social communication dataset consisting of 5 million users, finding that human interaction behavior dynamics exhibit rich complexities. We plot the

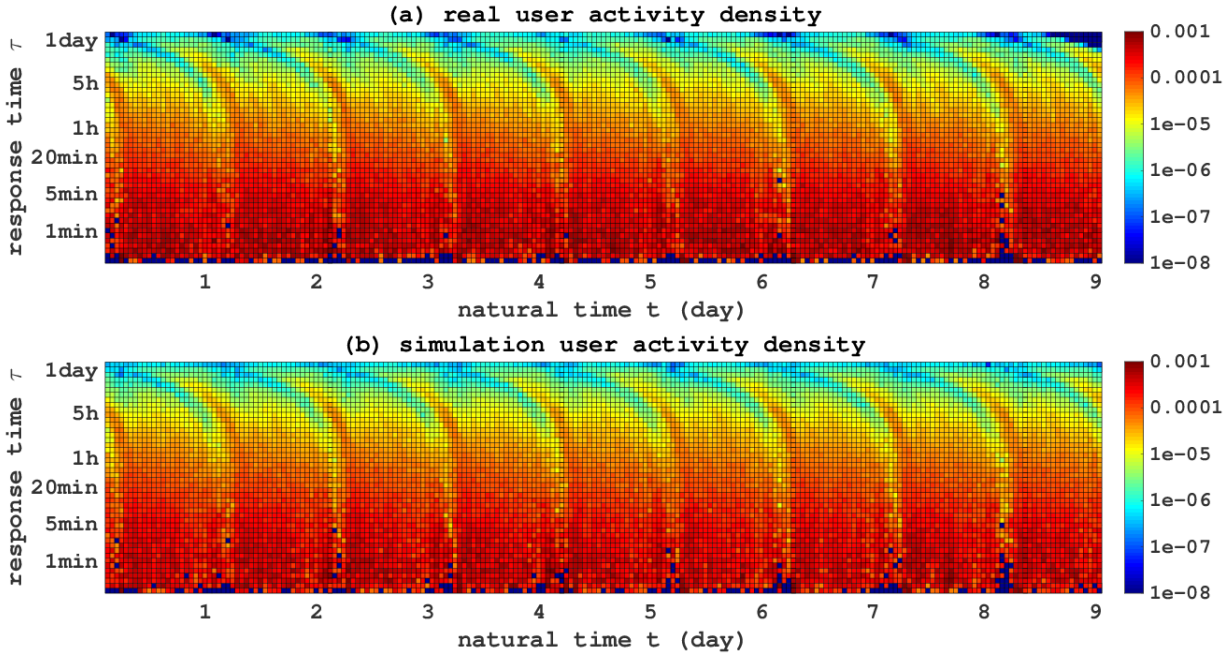

Figure 1: User activity densities of (a) an online information diffusion dataset described in section 4.1, (b) the simulation generated from our model.

response density functions $P_t(\tau)$ of user activities versus different starting time t in natural time scale in Figure 1(a). It is obvious that the densities change over response time, while the response time distribution also change over the natural time dimension following an obvious circadian rhythm. This suggests that the social behavior dynamics are temporally heterogeneous in nature. Although there have been some existing models to separately address the regularities in response time [35] and circadian rhythm of human behaviors [19], none of them can jointly model these two components, leading to non-trivial bias and error in understanding and predicting human behaviors.

In this paper, we propose a temporally Heterogeneous Survival framework where the regularities in response time dimension and natural time dimension are organically integrated. Our proposed model has the following advantages:

- Unification power: it is able to model the problems in both natural time scenarios and response time scenarios. As the model is designed under the probabilistic framework, it can be easily solved by Maximum Likelihood Estimation.
- Interpretability: All parameters have clear physical meanings. This is helpful for insightful understanding on social dynamics.
- Usefulness: We apply the model in two online social communication datasets. The learnt parameters and discovered statistical regularities lead to multiple potential applications.
- Accuracy: as shown in Figure 1(b), our model can successfully regenerate the interaction patterns in a social communication dataset. Also, extensive experiments are conducted to demonstrate the effectiveness of the model.

The rest of the paper is organized as follows: in Section 2, we give a survey on the related work. Section 3 presents a general framework of Heterogeneous Survival Model. Based on the observation of the dataset, we design a survival function modeling the social communication dynamics in Section 4. We evaluate our method and report the experimental result in Section 5. Last, we conclude our paper in Section 6.

2 RELATED WORKS

Research works in social dynamics aim to understand and model the information and knowledge spreading dynamics over social systems. Recently, much effort has been made towards this field thanks to the increasing availability of large-scale datasets, leading to the discovery of a number of generic mechanisms and ingredients governing social dynamics across various domains.

A lot of works in network science community aim to model diffusion dynamics through epidemic models in a continuous time basis, leading to critical theoretical advances such as the finding of absence of the epidemic threshold for inhomogeneous networks [2, 12, 22, 31, 34].

Complementarily, studies in the data mining community are interested in extracting and mining information from the real-world datasets. For example, a large portion of works aimed to cluster diffusion dynamics by distinct user interests and other human activities [1, 3–5, 10, 13, 14, 16–18, 21, 24–27, 30, 33]. However, most data mining approaches focus on

seeking relevant features and metrics through exploratory data analysis rather than offering generative models in a dynamic fashion.

More recently, new approach emerges by fusing the techniques developed in both areas, aiming to model the diffusion dynamics through continuous time Markov processes [6, 20]. Subsequent works based on the survival theory have relaxed the Markov condition, aiming to capture the non-Poisson nature of the observed waiting time distribution [9, 23, 28, 29, 35]. However, such kind of works can not deal with the key factor that the diffusion dynamics is influenced by the circadian rhythm [11, 32]. Although a few works have tried to solve this situation by heterogeneous poisson process [19], no models can consider both properties existing in diffusion dynamics.

3 TEMPORALLY HETEROGENEOUS SURVIVAL FRAMEWORK

In order to consider the complex waiting time pattern and inhomogeneous user activity in human behavior, we design a Temporally Heterogeneous Survival Framework depending on natural time variable t and response time variable τ. It tries to answer the following questions: if an event starts at t, what is the probability that a response event occurs after a certain duration τ? Of those that did not occur (at $t+\tau$), at what rate will they happen?

Based on the target of our framework, we propose three metrics as below:

- $f_t(\tau)$: probability density function, to record the probability that an event starts at t while the response duration is τ.
- $S_t(\tau)$: survival function, the complement of the c.d.f, which gives the probability of the response event did not happen before $t+\tau$.
- $\lambda_t(\tau)$: hazard function, or intensity function, the conditional probability that the event will occur in $t+\tau$ if it did not happened before $t+\tau$.

Given one of these metrics, the other two metrics can also be determined by the following equations:

$$S_t(\tau) = Pr(T \geq \tau) = \int_\tau^\infty f_t(\tau')d\tau' \tag{1}$$

$$f_t(\tau) = -\frac{\partial S_t(\tau)}{\partial \tau} \tag{2}$$

$$\lambda_t(\tau) = \frac{f_t(\tau)}{S_t(\tau)} = -\frac{\partial S_t(\tau)}{\partial \tau}\frac{1}{S_t(\tau)} = \frac{f_t(\tau)}{\int_\tau^\infty f_t(\tau')d\tau'} \tag{3}$$

$$S_t(\tau) = e^{-\int_0^\tau \lambda_t(\tau')d\tau'} \tag{4}$$

$$f_t(\tau) = \lambda_t(\tau)e^{-\int_0^\tau \lambda_t(\tau')d\tau'} \tag{5}$$

For ease of modeling, we often try to model the hazard function when the survival process is complex. It usually has succinct formulation comparing with the other two metrics.

Next, we will give a formal definition of the communication data, to explain what kind of data our model can handle.

3.1 Communication data format

Given a sender u and a receiver v, the communication data of (u, v) is a joint set $\Omega_{u,v}$ that consists of all the communication records between u and v: $\Omega_{u,v} = \{(s_i, e_i)\}$. Here, A communication record (s, e) means one user (receiver) begins a communication at s, and gets the response from another user (sender) at e.

3.2 The Likelihood Function

Suppose that we have n communication records $\{(s_i, e_i)\}$ governed by a survival function $S_t(\tau|\theta)$ with associated density function $f_t(\tau|\theta)$ and hazard function $\lambda_t(\tau|\theta)$ under parameters θ. The likelihood function of communication records can then be written as follows:

$$\begin{aligned} L &= \Pi_{i=1}^n f_{s_i}(e_i - s_i|\theta) \\ &= \Pi_{i=1}^n \lambda_{s_i}(e_i - s_i|\theta)S_{s_i}(e_i - s_i|\theta) \end{aligned} \tag{6}$$

Taking the logarithm of the survival function, we obtain the log-likelihood function for the communication data:

$$\begin{aligned} l &= \log L \\ &= \sum_{i=1}^n (\log \lambda_{s_i}(e_i - s_i|\theta) + \log S_{s_i}(e_i - s_i|\theta)) \\ &= \sum_{i=1}^n \left(\log \lambda_{s_i}(e_i - s_i|\theta) - \int_0^{e_i - s_i} \lambda_{s_i}(x|\theta)dx\right) \end{aligned} \tag{7}$$

The survival parameters θ can then be measured by maximizing the log-likelihood function (Equation 7). Usually, we can use the L-BFGS Quasi-Newton Method [15] to solve this problem. We can first get the derivatives of all parameters with respect to λ, and then get the derivatives to the log-likelihood function by using Equation 8 and Chain Rule in calculus.

$$\begin{aligned} \frac{\partial l}{\partial \theta} &= \frac{\partial \sum_{i=1}^n \left(\log(\lambda_{s_i}(e_i - s_i|\theta)) - \int_0^{e_i - s_i} \lambda_{s_i}(t'|\theta)dt'\right)}{\partial \theta} \\ &= \sum_i \left(\frac{\partial \lambda_{s_i}(e_i - s_i|\theta)/\partial\theta}{\lambda_{s_i}(e_i - s_i|\theta)} - \int_0^{e_i - s_i} \frac{\partial \lambda_{s_i}(t'|\theta)}{\partial \theta}dt'\right) \end{aligned} \tag{8}$$

3.3 General Decomposition

Usually, it is hard to design a comprehensible hazard function due to the complex pattern in two-dimensional density function. Hence, an intuitive idea is to model the hazard function by adopting a divide-and-conquer approach: we can first design a stretching function $\omega(t)$, indicating the activeness of the user at different time t (Thereby, the stretching function will have a mean value of 1). Simultaneously, we may also design another respond function $G(\tau)$ for the purpose of modeling the intensity at different response time τ. After that, the hazard function can be finalized by a combination of these two functions:

$$\lambda_t(\tau) = \omega(t)G(\omega(t)\tau) \tag{9}$$

Given that it can include all previous studies as special cases, we consider such kind of decomposition very powerful. For example, when $\omega(t)$ remains the same value (equal to 1) at every point, the hazard function will degenerate to:

$$\lambda_t(\tau) = G(\tau) \tag{10}$$

leading to the same form of the homogeneous survival model.

On the other hand, when G is unchanged, the hazard function will be:

$$\lambda_t(\tau) = c_2\omega(t) \tag{11}$$

In this situation, our heterogeneous survival model will degenerate to the heterogeneous poisson model. Thereby, it is clear that our model gains the properties from both homogeneous survival model and heterogeneous poisson model.

4 MODELING RESPONSE DYNAMICS IN SOCIAL SYSTEM

In this section, we will first give an introduction to the dataset we used in our paper. After that, we will give a solution on modeling the patterns we discovered in the dataset.

4.1 Dataset Description

4.1.1 Online information diffusion data. [1]

This is an online information diffusion dataset [29] from Tencent Weibo[2], a Twitter-style social platform in China. It includes all cascades with at least 5 tweets generated in the 10 days between Nov 15 and Nov 25 2011. For each tweet, there is a triad $< u, t, r >$ to respectively represent the sender of tweet, sending time of tweet, and the user it reply to. Therefore, it is easy to obtain a communication record (t', t) by getting the tweet time t' sent by r for each tweet. In total, there are 0.46 million cascades in the dataset.

4.1.2 Email data. This dataset contains a 83-days email communication records between users in an university mail server [7]. Each record in the dataset is constituted by a triad, e.g., <time t, sender r, receiver s >. For each record $< t, r, s >$, we find the latest record $< t', s, r >$ sent before t as a communication record (t', t). In order to remove the communication record triggered by the automatic response in the mail system, we omit the communication record (t', t) whose response duration is less than 10 second, e.g., $t - t' \leq 10$.

4.2 Modeling Inhomogeneity in Daily activity

We next examine the daily activity factor, which is not considered in most previous studies. We count the number of messages per hour every day in the information diffusion dataset, and illustrate the result at Figure 2. We can see that the user activities have a clear circadian patterns: people are more active at daytime and less active at night.

[1]The dataset is complete and publicly available at http://www.thumedia.org

[2]http://t.qq.com

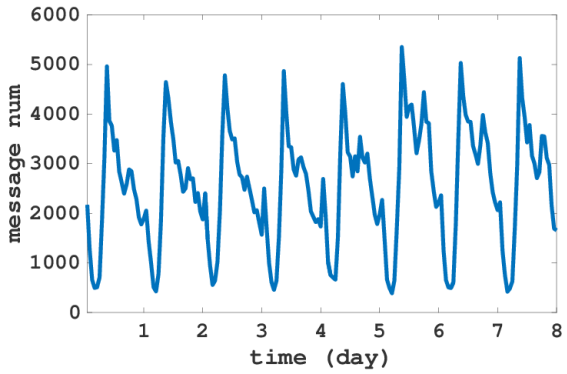

Figure 2: The number of messages per hour every day in the information diffusion dataset.

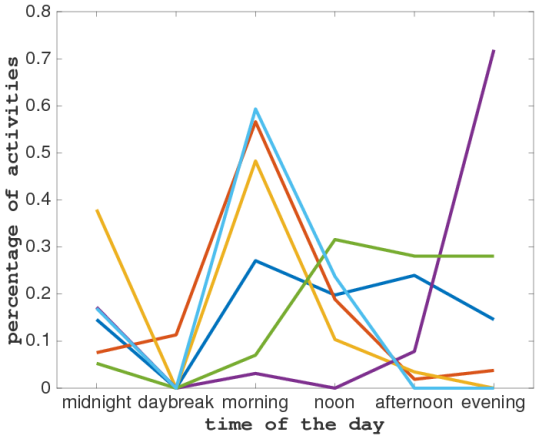

Figure 3: diverse user activities in a social system

In order to eliminate the distinct survival effect caused by the rise and fall pattern of daily activities, several previous studies try to quantify the dynamics by event time instead of natural time [8, 28, 29, 35] [3]. The idea is very similar with our design purpose of $\omega(t)$. However, it has the following drawbacks:

- It can not provide insights on how people behave in natural time.
- The design is not precise. By plotting the daily activity values of six different users in Figure 3, we can find that the user daily activities are very dissimilar[4].

Hence, the idea of event time is not suitable for such kind of dynamics.

[3]redefine the time by the number of messages users post on the social system

[4] Here, we divide the day into 6 periods, including:
- daybreak: 2:00-6:00
- morning: 6:00-10:00
- noon: 10:00-14:00
- afternoon: 14:00-18:00
- evening: 18:00-22:00
- midnight: 22:00-2:00(nextday)

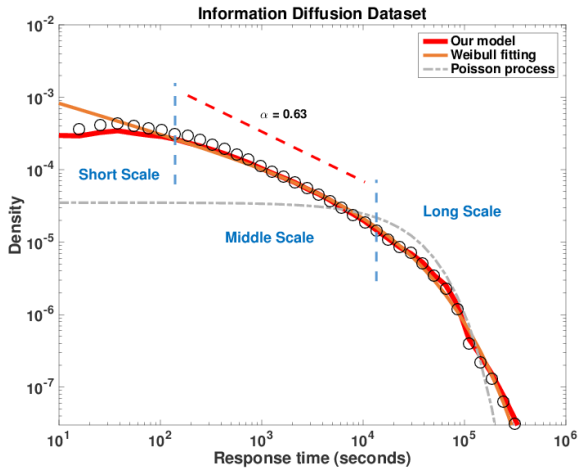

Figure 4: The probability density function of the time interval between user communications in the information diffusion dataset.

In order to model the circadian activities of a user, we provide a Periodic Gaussian Mixture Model as below:

$$\omega(t) = \sum_{i=0}^{k} \phi_i \int_{t'=-\infty}^{\infty} III_T(t') g(t'|\mu_i, \sigma_i^2) dt' \tag{12}$$

Here, $III_T(t)$ is a Dirac comb[5] with the periodic parameter T, and $g(t'|\mu, \sigma^2)$ denotes the density of a normal distribution. The periodic parameter T can be used in various domains. For example, it will be a daily periodic function when $T = 86400s$, and a weekly periodic function when $T = 7days$. In our model, we choose to use a daily periodic function.

4.3 Modeling inhomogeneity in response time

We further analyze the inhomogeneity in the response time dimension. we estimate the probability density function of the time interval between user communications, and illustrate it in Figure 4. As we can see, the response time ranges over five orders of magnitudes. Basically, it can be classified into three parts. For the middle scale, it can be well approximated by a power-law function. However, the power-law characteristic can not capture the pattern in shorter and longer time scales. The sharp cutoff at large time scale indicates a clear exponential tail. Nevertheless, it is very dissimilar between real data and the homogeneous poisson process.

Thereby, we proposed the following Weibull based generic intensity function to incorporate different activity patterns in heterogenous dimension:

$$G(\tau) = \lambda_0 \tau^{-\alpha} H(\tau - \beta) \tag{13}$$

where the parameters $\{\lambda_0, \alpha, \beta\}$ captures the following different aspects in human activity patterns:

[5] http://en.wikipedia.org/wiki/Dirac_comb

- λ_0 is the scale parameter of the distribution. It mainly controls the average duration of the response. The smaller the scale parameter, the more spread out the distribution.
- $\alpha(\alpha < 1)$ is the shape parameter which describes the relationship between the intensity rate and response time:
  - $\alpha > 0$: in this situation, the intensity rate is decreasing over time. This happens when most of the responses occurs in the very early stage, or there is a significant "infant mortality". Hence, it is very important to consider the communications with such patterns when we want to maximize the influence.
  - $\alpha = 0$: in this situation, the intensity rate is independent of the response duration. It is proportional to the daily activity value, and the whole process will reduce to a heterogeneous poisson process.
  - $\alpha < 0$: this means the intensity rate increases over time. In this situation, the responses are more likely to occur as time goes on, which means it is an "aging" process.
- β is the location parameter, which determines the "location" or the shift of the distribution. The $H(x)$ function used in the formula is a Heaviside step function, which is a discontinuous function whose value is 0 when $x < 0$ and 1 when $x > 0$. In social dynamics scenario, the location parameter can indicate the consideration duration in the response of a user.

4.4 Parameter Optimization

Based on the discovery and the proposed function in Equation 12 and Equation 13, the hazard function can be obtained using Equation 9 as:

$$\lambda_t(\tau|\theta) = \frac{\lambda_0 \omega(t)}{(\omega(t)\tau)^{\alpha}} H(\omega(t)\tau - \beta) \tag{14}$$

For ease of computation, we replace the Heaviside step function by a logistic function (with a fix parameter $\phi = 0.001$), and finalize the hazard function as:

$$\lambda_t(\tau|\theta) = \frac{\lambda_0 \omega(t)}{(\omega(t)\tau)^{\alpha}} \left(1 + e^{-\frac{\omega(t)\tau - \beta}{\phi}}\right)^{-1} \tag{15}$$

We only need to get the gradients of every parameters with respect to $\lambda_t(\tau)$, and then using the strategy proposed in Section 3.2 to get the modeling parameters $\theta = \{\lambda_0, \alpha, \omega(t), \beta\}$:

$$\frac{\partial \lambda_t(\tau)}{\partial \lambda_0} = \frac{\omega(t)}{(\omega(t)\tau)^{\alpha}} \left(1 + e^{-\frac{\omega(t)\tau - \beta}{\phi}}\right)^{-1} \tag{16}$$

$$\frac{\partial \lambda_t(\tau)}{\partial \alpha} = -\log(\omega(t)\tau) \frac{\lambda_0 \omega(t)}{(\omega(t)\tau)^{\alpha}} \left(1 + e^{-\frac{\omega(t)\tau - \beta}{\phi}}\right)^{-1} \tag{17}$$

$$\frac{\partial \lambda_t(\tau)}{\partial \omega(t)} = (1-\alpha)\frac{\lambda_0}{(\omega(t)\tau)^\alpha}\left(1+e^{-\frac{\omega(t)\tau-\beta}{\phi}}\right)^{-1} + \frac{\lambda_0\omega(t)}{(\omega(t)\tau)^\alpha}\frac{\tau}{\phi}\left(1+e^{-\frac{\omega(t)\tau-\beta}{\phi}}\right)^{-2} e^{-\frac{\omega(t)\tau-\beta}{\phi}} \quad (18)$$

$$\frac{\partial \omega(t)}{\partial \phi_i} = \sum_{i=0}^{k}\int_{t'=-\infty}^{\infty} III_T(t')g(t'|\mu_i,\sigma_i^2)dt' \quad (19)$$

$$\frac{\partial \omega(t)}{\partial \mu_i} = \sum_{i=0}^{k}\phi_i\int_{t'=-\infty}^{\infty} III_T(t')\frac{\partial g(t'|\mu_i,\sigma_i^2)}{\partial \mu_i}dt' \quad (20)$$

$$\frac{\partial \omega(t)}{\partial \sigma_i} = \sum_{i=0}^{k}\phi_i\int_{t'=-\infty}^{\infty} III_T(t')\frac{\partial g(t'|\mu_i,\sigma_i^2)}{\partial \sigma_i}dt' \quad (21)$$

$$\frac{\partial \lambda_t(\tau)}{\partial \beta} = -\frac{\lambda_0\omega(t)}{(\omega(t)\tau)^\alpha}\frac{1}{\phi}\left(1+e^{-\frac{\omega(t)\tau-\beta}{\phi}}\right)^{-2} e^{-\frac{\omega(t)\tau-\beta}{\phi}} \quad (22)$$

For the purpose of finding a good region of parameter space, we set the initial value of parameters using some prior knowledge. Meanwhile, in order to avoid the accuracy error, we divide the parameters into three groups, and optimize these parameters iteratively until reaching convergence.

The overall algorithm is presented in Algorithm 1.

Algorithm 1 Parameter Optimization

Input:
Communication data of two users $CD(u,v)$.
Output:
Communication parameters $\{\lambda_0, \alpha, \beta, \omega(t)\}$.

```
it = 0
sum_duration = 0;                                   ▷ Initialization.
min_duration = ∞;
for i=1 to size(CD(u,v)) do
    sum_duration = sum_duration + CD(u,v)(i).e - CD(u,v)(i).s;
    if CD(u,v)(i).e - CD(u,v)(i).s ¡ min_duration then
        min_duration = CD(u,v)(i).e - CD(u,v)(i).s
    end if
end for
average_duration = sum_duration / size(CD(u, v));
λ0(t)^[0] = 1/ average_duration
α^[0] = 0.9
get ω(t)^[0] using KMEANS with all e_i in CD(u,v)
β^[0] = 0.5 min_duration                            ▷ End initialization.
repeat
    it = it + 1
    λ0^[it], α^[it] = argmax_{λ0,α} l(λ0, α, β^[it-1], ω(t)^[it-1])
    β^[it] = argmax_β l(λ0^[it], α^[it], β, ω(t)^[it-1])
    ω(t)^[it] = argmax_{ω(t)} l(λ0^[it], α^[it], β^[it], ω(t))
until Convergence
return θ = {λ0^[it], α^[it], β^[it], ω(t)^[it]}
```

5 EXPERIMENTS

In this section, we conduct experiments on two datasets introduced in Section 4.1. We will first show how well our model matches the real world data, and then analyze the distribution of all the parameters to illustrate the distinction between different social systems. Furthermore, we will give suggestions on how to maximize the effect of information spreading among diverse groups of people.

5.1 Baselines and Evaluation Metrics

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

- Weibull distribution: it gives a distribution which has been controlled by two parameters and k in its scale and shape. The failure rate is proportional to a power of time.
- Log Normal distribution: it is a continuous probability distribution of a random variable whose logarithm is normally distributed. Hence, $Y = \ln(X)$ has a normal distribution if X is log-normally distributed.
- Pareto distribution: the Pareto distribution is a power law probability distribution that is used in many types of observable phenomena.
- Heterogeneous Poisson Point Process: the Poisson point process is one of the most used point processes. When the intensity parameter λ varies over time, it is called the Heterogeneous Poisson Point Process.

As some previous studies use event time to eliminate the natural time user activity effect, we examine the efficiency for the homogeneous survival based model in two ways:

- We model the communication dynamics directly; or
- At the outset, we transfer the natural time communication data to event time communication data and then learn the model and regenerate the data under event time scenario. In the end, we transfer the event time regeneration back to the natural time domain.

We use KolmogorovSmirnov test to evaluate the performance. It is one of the most useful methods for comparing two samples. It tries to quantify a distance between the empirical cumulative distribution functions of two samples by the following method:

$$ksstat = sup_x|F_1(x) - F_2(x)| \quad (23)$$

where the empirical cumulative distribution function for a sample is defined as below:

$$G(x) = \frac{1}{n}\sum_{1}^{n} I_{[-\inf,x]}(X_i) \quad (24)$$

$$I_{[-\inf,x]}(X_i) = (X_i \le x)$$

The lower the $ksstat$ is, the more similar the two samples will be. Based on the $ksstat$ value, we further calculate the $p-value$ of the test case, and calculate out the $ksrate$ value as the percentage of test cases which pass the $p-value$ test at the default 5% significance level. In so doing, the larger the average value of $ksrate$ is, the more accurate the model will achieve.

5.2 Effectiveness

We first make a comparison on the information diffusion dataset. We try to evaluate the methods in three dimensions:

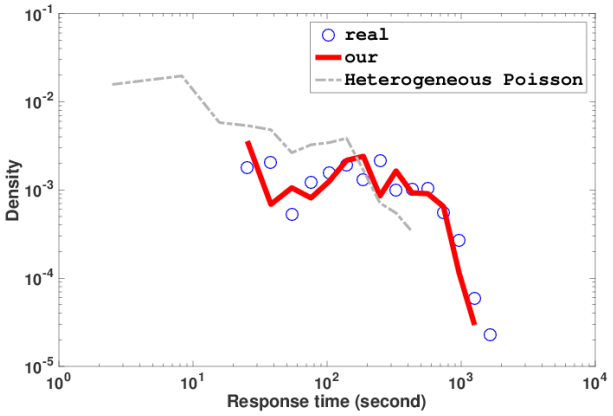

Figure 5: Response time distribution of a test case in the dataset.

- **Response time duration KS-test**: to test whether the response time distribution in the simulation is similar with the real case.
- **Daily activities KS-test**: to test whether the activities we generated in different time period of the day are similar with the real case.
- **Accuracy**: the probability that the test case passes both the two tests above.

Table 1 shows the experimental result. Although our method does not give the best performance in the KS-test of the response time, we beat the rest of methods in all the other metrics. 77.1% cases of our model successfully pass both the **Response time duration KS-test** and the **Daily activities KS-test**, while the other methods can only achieve 56.4%. The improvement is about 36.6%.

The Heterogeneous Poisson Point Process is the closest method compared at the accuracy test. Comparing to the other baselines, it has advantage at the Daily Activities Dimension. However, as we expected, it can not well capture the response time distribution as shown in Figure 5.

For the rest of the baselines, the survival based methods give a similar performance in the **Response time duration KS-test**. However, all of these methods cannot well capture the user activities in natural time scenario. Here, we plot the number of the response messages per hour everyday in Figure 6. Unlike our model, these methods will generate much more response messages in late night and much less messages in the daytime comparing to the real data. The experiment result improves if we conduct transformation between natural time and event time before and after modeling. However, there is still a huge gap between the methods using event time transformation and our model.

We use another statistic test to further demonstrate the advantage of our model: for one test case, we first classify the communication data (and the simulation communication data) into several sub−data by the starting time of the communication record, then apply the KS-test to each sub-data. The performances for all the methods are shown in Figure 7. It can be seen that the proposed method significantly outperforms other baselines in all time period.

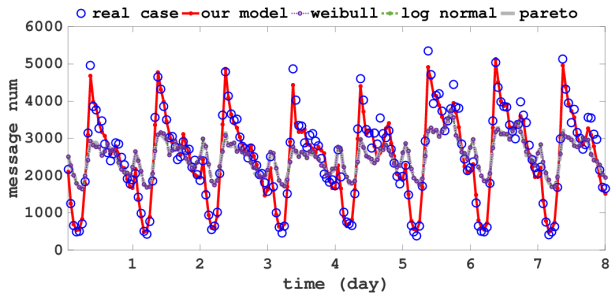

Figure 6: The comparison in the number of response messages generated by the models.

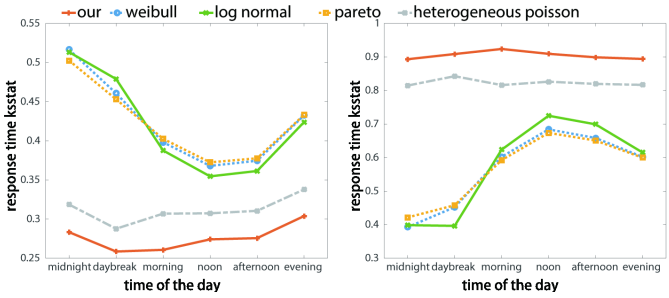

Figure 7: *ksrate* and its *ksstat* of information diffusion data clustered by time

We also carry out the same experiment on the email dataset, and present the result on table 2. Similarly, our model shows no advantage if we only take a look at the *ksstat* on the response time dimension. However, our method has higher performance on *ksstat* in daily activity and overall *accuracy* comparing to all the other methods. The *ksrate* and *ksstat* in daily activity dimension of all the other survival based methods get a better value comparing to their performance in the information diffusion dataset. A possible reason is that people tend to receive and respond emails in some critical time point of the day, while their time in the social network are more diverse. Therefore, our model are more suitable for the heavy users in the social network.

5.3 Model Parameter Analysis

We further analyze the parameters of our model from two perspectives by their physical meanings. On one hand, we try to find when the users tend to use the social system from their daily activity parameters $\omega(t)$. On the other hand, we analyze what kind of behaviors they tend to do based on their respond function parameters.

5.3.1 Daily Activities. According to the daily activity parameters $\omega(t)$ learnt from our model, the users in the information diffusion dataset can be divided into 7 categories by adopting the k-means algorithm. The result is illustrated on Figure 8. We can further combine them into 4 groups with clear meanings:

- Heavy users: the users in category 5 stay active all day in the social system. In our information dataset, there are 39.49% users having this attributes.

Table 1: Fitting result for the information diffusion dataset. The ksrate is the pass rate at the default 5% significance level.

Model	our	Weibull		Log Normal		Pareto		Heterogeneous
		real	event	real	event	real	event	Poisson
Response ksstat	0.1597	0.1555	0.1612	0.1741	0.1731	0.1581	0.1673	0.2021
Response ksrate	87.23%	88.56%	87.23%	80.88%	81.64%	87.56%	84.11%	71.43%
Daily ksstat	0.1440	0.3532	0.3190	0.3520	0.3249	0.3526	0.3189	0.1907
Daily ksrate	84.49%	28.91%	33.03%	30.11%	32.61%	30.38%	33.99%	73.52%
Accuracy (both pass)	**77.06**%	26.74%	30.72%	27.91%	30.30%	28.79%	32.00%	56.40%

Table 2: Fitting result for email dataset. The ksrate is the pass rate at the default 5% significance level.

Model	our	Weibull		Log Normal		Pareto		Heterogeneous
		real	event	real	event	real	event	Poisson
Response ksstat	0.2150	0.1813	0.2022	0.1748	0.1945	0.1685	0.1897	0.2797
Response ksrate	79.07%	94.57%	88.37%	96.12%	86.82%	96.90%	93.80%	40.31%
Daily ksstat	0.1440	0.2527	0.2230	0.2420	0.2188	0.2385	0.2189	0.1570
Daily ksrate	100%	65.89%	76.74%	68.99%	83.72%	73.64%	78.29%	100%
Accuracy (both pass)	**79.07**%	63.57%	67.44%	68.22%	72.09%	73.64%	73.64%	40.31%

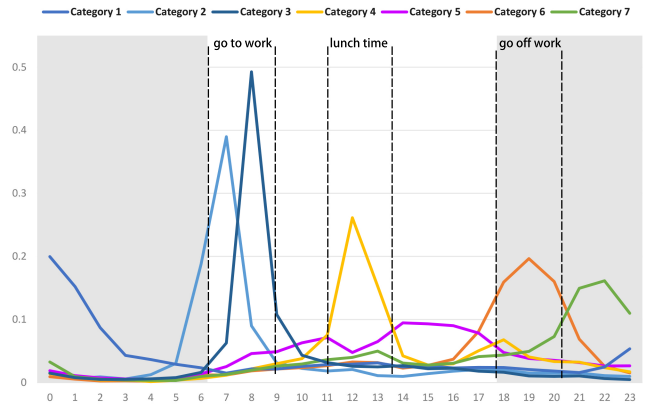

Figure 8: The clustering result of user daily activities. The white area means the day time, while the grey area means the evening time.

- For the purpose of entertainment: some of the users like to use the social system between dinner and bed time (as shown in category 7). The users with the habit will be labeled as "For the purpose of entertainment". In all, there are 13.52% users with this label.
- Using as a leisure tool: 21.44% people tend to use the service as a leisure tool before sleep, leading to a peak at 12 : 00 p.m. in category 4, and another peak at 0 : 00 a.m. in category 1.
- Using on the way: there are 25.55% people who tend to use the social network service during the commuter time. We can observe obvious peaks at $7:00-9:00$ a.m. in category $2, 3$ and $18:00-20:00$ p.m. in category 6.

We can find similar results in the clustering of the email dataset. Given the length of the paper, we did not show the clustering result of the email dataset.

5.3.2 Respond Activities. Next, we count the parameter distributions of the respond function parameters in both datasets, and present the result in Figure 9 and Figure 10. From the figure, we can get the following findings:

- λ_0: the distribution of λ_0 follows a log normal distribution in both datasets. The mean value of λ_0 in the cascading dataset is around e^{-10}, while most λ_0 in the email dataset locate at e^{-9}.
- α: it follows a normal distribution with mean value 0.008 in cascading dataset and 0.1206 in email dataset.

  In the information diffusion dataset, 10.8% test cases has a α parameter whose absolute value is less or equal than 0.01. In such situation, it is a heterogeneous poisson process dependent on daily activities (we have already explained in section 4.3), which partially proves the findings in [19]. There are still 36.4% test cases whose $\alpha > 0.01$, and 52.8% test cases whose $\alpha < -0.01$, indicating that the failure rate of these processes is dependent on the response durations. Meanwhile, most test cases in the email dataset have a positive α value (91.4%), showing that the failure rate is decreasing over time. If we want to get response from these people (with $\alpha > 0$), then we need to seriously consider the time start the communication.
- β: the distribution of β is different with two datasets. In the email dataset, it is a typical log normal distribution with a single peak near $e^5 s \sim 2min$. In the information diffusion dataset, it is a bimodal distribution with one peak near 0 and the other near $1min$, showing that two

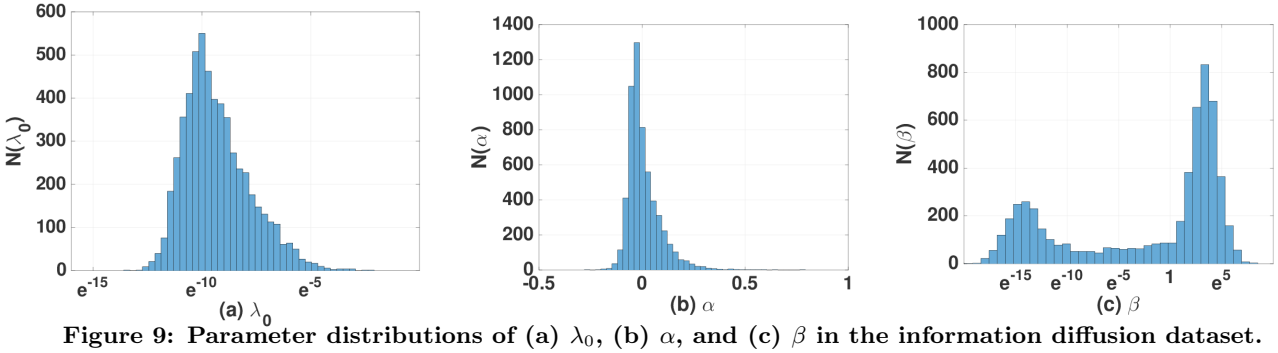

Figure 9: Parameter distributions of (a) λ_0, (b) α, and (c) β in the information diffusion dataset.

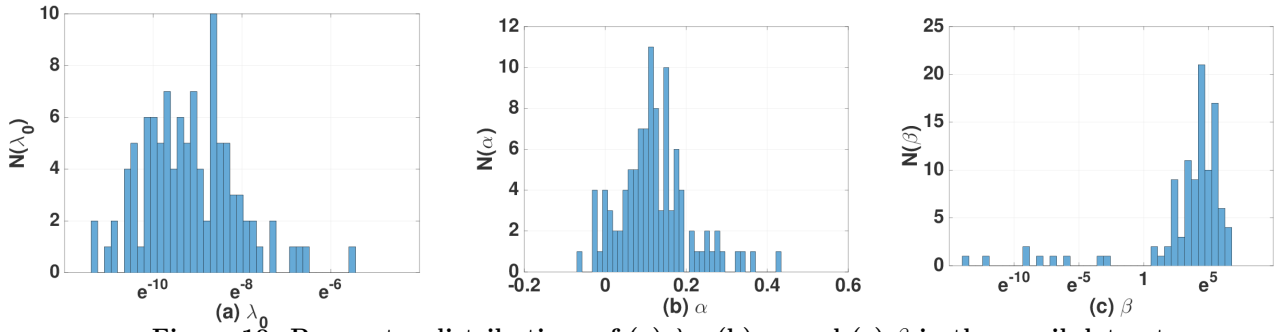

Figure 10: Parameter distributions of (a) λ_0, (b) α, and (c) β in the email dataset.

different behaviors exist in the social network system: some people like to make decisions very fast, while the others tend to think for a while before their response. Hence, we need to use more eye-catching words to attract the people having smaller β on one hand, while, on the other hand, concentrate more on the substantive content for those with higher β.

6 CONCLUSIONS

In this paper, we study the problem of social behavior dynamics modeling. In order to solve this problem, we propose a temporally Heterogeneous Survival framework, and give a novel method that models the communication intensity rate between two people based on the observation from an information diffusion dataset. Our proposed model has the following advantages:

- Unification power: it is able to model the problems in both natural time scenarios and response time scenarios. As the model is designed under the probabilistic framework, it can be easily solved by Maximum Likelihood Estimation.
- Interpretability: All parameters have clear physical meanings. This is helpful for insightful understanding on social dynamics.
- Usefulness: We apply the model in two online social communication datasets. The learnt parameters and discovered statistical regularities lead to multiple potential applications.
- Accuracy: our model can successfully regenerate the interaction patterns in a social communication dataset. Also, extensive experiments are conducted to demonstrate the effectiveness of the model.

7 ACKNOWLEDGEMENT

This work was supported by National Program on Key Basic Research Project, No. 2015CB352300; National Natural Science Foundation of China, No. 61370022, No. 61521002, No. 61531006 and No. 61210008. Thanks for the research fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology. Song was partly supported by the National Science Foundation (IBSS-L- 1620294), and by a Convergence Grant from the College of Arts & Sciences, University of Miami.

REFERENCES

[1] Yoav Artzi, Patrick Pantel, and Michael Gamon. 2012. Predicting responses to microblog posts. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 602–606.

[2] Marián Boguná, Romualdo Pastor-Satorras, and Alessandro Vespignani. 2003. Absence of epidemic threshold in scale-free networks with degree correlations. *Physical review letters* 90, 2 (2003), 028701.

[3] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *Proceedings of the 23rd international conference on World wide web*. ACM, 925–936.

[4] Peng Cui, Fei Wang, Shaowei Liu, Mingdong Ou, Shiqiang Yang, and Lifeng Sun. 2011. Who should share what?: item-level social influence prediction for users and posts ranking. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 185–194.

[5] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 307–318.

[6] Nan Du, Le Song, Hyenkyun Woo, and Hongyuan Zha. 2013. Uncover topic-sensitive information diffusion networks. In *Proceedings of the sixteenth international conference on artificial*

intelligence and statistics. 229–237.

[7] Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi. 2004. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America* 101, 40 (2004), 14333–14337.

[8] Shuai Gao, Jun Ma, and Zhumin Chen. 2015. Modeling and Predicting Retweeting Dynamics on Microblogging Platforms. In *Eighth ACM International Conference on Web Search and Data Mining*. 107–116.

[9] Manuel Gomez-Rodriguez, Jure Leskovec, and Bernhard Schölkopf. 2013. Modeling Information Propagation with Survival Theory.. In *ICML (3)*. 666–674.

[10] Lei Guo, Enhua Tan, Songqing Chen, Xiaodong Zhang, and Yihong Eric Zhao. 2009. Analyzing patterns of user content generation in online social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 369–378.

[11] M. Karimi, E. Tavakoli, M. Farajtabar, L. Song, and M. Gomez-Rodriguez. 2016. Smart Broadcasting: Do you want to be seen? (2016).

[12] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. 2010. Identification of influential spreaders in complex networks. *Nature physics* 6, 11 (2010), 888–893.

[13] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*. ACM, 591–600.

[14] Himabindu Lakkaraju, Julian J McAuley, and Jure Leskovec. 2013. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. *ICWSM* 1, 2 (2013), 3.

[15] Dong C Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical programming* 45, 1 (1989), 503–528.

[16] Zhunchen Luo, Miles Osborne, Jintao Tang, and Ting Wang. 2013. Who will retweet me?: finding retweeters in twitter. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 869–872.

[17] Zongyang Ma, Aixin Sun, and Gao Cong. 2012. Will this# hashtag be popular tomorrow?. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1173–1174.

[18] Zongyang Ma, Aixin Sun, and Gao Cong. 2013. On predicting the popularity of newly emerging hashtags in twitter. *Journal of the American Society for Information Science and Technology* 64, 7 (2013), 1399–1410.

[19] R Dean Malmgren, Daniel B Stouffer, Adilson E Motter, and Luís AN Amaral. 2008. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences* 105, 47 (2008), 18153–18158.

[20] Seth Myers and Jure Leskovec. 2010. On the convexity of latent social network inference. In *Advances in Neural Information Processing Systems*. 1741–1749.

[21] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd International Web Science Conference*. ACM, 8.

[22] Romualdo Pastor-Satorras and Alessandro Vespignani. 2001. Epidemic spreading in scale-free networks. *Physical review letters* 86, 14 (2001), 3200.

[23] Manuel Gomez Rodriguez, David Balduzzi, and Bernhard Schölkopf. 2011. Uncovering the temporal dynamics of diffusion networks. *arXiv preprint arXiv:1105.0697* (2011).

[24] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*. IEEE, 177–184.

[25] Oren Tsur and Ari Rappoport. 2012. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 643–652.

[26] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, 705–714.

[27] Louis Yu, Sitaram Asur, and Bernardo A Huberman. 2011. What trends in Chinese social media. *arXiv preprint arXiv:1107.3522* (2011).

[28] Linyun Yu, Peng Cui, Fei Wang, Chaoming Song, and Shiqiang Yang. 2015. From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics. In *Data mining (ICDM), 2015 IEEE international conference on*. IEEE, 559–568.

[29] Linyun Yu, Peng Cui, Fei Wang, Chaoming Song, and Shiqiang Yang. 2016. Uncovering and predicting the dynamic process of information cascades with survival model. *Knowledge and Information Systems* (2016), 1–27.

[30] Tauhid R Zaman, Ralf Herbrich, Jurgen Van Gael, and David Stern. 2010. Predicting information spreading in twitter. In *Workshop on computational social science and the wisdom of crowds, nips*, Vol. 104. Citeseer, 17599–601.

[31] Chengxi Zang, Peng Cui, and Christos Faloutsos. 2016. Beyond Sigmoids: The NetTide Model for Social Network Growth, and Its Applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015–2024.

[32] A. Zarezade, U. Upadhyay, H. Rabiee, and M. Gomez-Rodriguez. 2017. RedQueen: An Online Algorithm for Smart Broadcasting in Social Networks. In *WSDM '17: Proceedings of the 10th ACM International Conference on Web Search and Data Mining*.

[33] Qi Zhang, Yeyun Gong, Ya Guo, and Xuanjing Huang. 2015. Retweet Behavior Prediction Using Hierarchical Dirichlet Process.. In *AAAI*. 403–409.

[34] Tianyang Zhang, Peng Cui, Christos Faloutsos, Yunfei Lu, Hao Ye, Wenwu Zhu, and Shiqiang Yang. 2016. Come-and-Go Patterns of Group Evolution: A Dynamic Model. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1355–1364.

[35] Tianyang Zhang, Peng Cui, Chaoming Song, Wenwu Zhu, and Shiqiang Yang. 2016. A multiscale survival process for modeling human activity patterns. *PloS one* 11, 3 (2016), e0151473.