Modeling the Dynamics of WeChat Social System

Chengxi Zang
CS Department, Tsinghua University
chengxi.zang@gmail.com

Peng Cui
CS Department, Tsinghua University
cuip@tsinghua.edu.cn

Linyun Yu
CS Department, Tsinghua University
yuly12@mails.tsinghua.edu.cn

Tianyang Zhang
CS Department, Tsinghua University
zhangty09@foxmail.com

Wenwu Zhu
CS Department, Tsinghua University
wwzhu@tsinghua.edu.cn

Hao Ye
Tencent Corp., Shenzhen, China
dariaye@tencent.com

ABSTRACT

The phenomena in social systems, just like many laws of nature, are of dynamic origins. For the first time, we systematically summarize our experiences in dealing with dynamic phenomena/data in WeChat social system, which is one of the largest, most complex social systems in the real world. A framework is proposed, which covers the pipeline of dealing with complex data, complex phenomena, complex mechanisms, a spectrum of methodologies, and various applications. By following our framework, we discuss four main aspects of dynamics in WeChat social system, including human dynamics, network dynamics, group dynamics and cascade dynamics. We abstract the data structures from the digital logs, investigate the phenomena we observed from the data, explain the possible mechanisms governing the phenomena, and give models we design according to our methodology spectrum to answer various "WHEN" related questions. We illustrate the "research network" underlying our framework when applied to four dynamic scenarios. Typical applications are summarized and illustrated. The framework we give could potentially guide future researches on social dynamics related problems, sheding light on modeling and designing data structures, explaining observed phenomena, inferring possible mechanisms, choosing and designing dynamic models, and pinpointing applications in the real-world social systems.

KEYWORDS

WeChat; Social Dynamics; Human Dynamics; Network Dynamics; Group Dynamics; Cascade Dynamics;

1 INTRODUCTION

What is WeChat? Words fall short of describing what WeChat really is. Originally, it appeared as a communication tool. Step by step, it has become the largest (online) social network in China, with 963 million monthly active users in the second quarter of 2017. Through WeChat, people communicate with each other via various messaging methods, such as text message, voice message, group chat, video call, etc. Now, WeChat has gone far beyond a communication tool; it comes to be a lifestyle for many. People can share their wonderful lives in the "moments" – a social feed function which allows users to post images, texts, videos, music, etc., to receive comments and likes, and to repost the information to others. Meanwhile, with 600 million active mobile payment users, it offers the largest payment service in China, covering the features like digital wallets, balance accounts, red packet, lucky money, cash transfer, financial investment, etc. People pay through WeChat in restaurants, supermarkets or even vendors by scanning QR codes rather than using cash or credit cards. Moreover, WeChat provides official accounts, city services, mini programs, WeChat index, news feed, search, and many more functions. With the convenient and personalized services provided by WeChat, people now exchange their WeChat accounts rather than the business cards when making friends or developing business. It's a super app, an ecosystem, an app for "everything" in China and beyond.

What are the dynamics in WeChat? The WeChat social system is constantly changing over time in all the scenarios described above. When we text each other, add a new friend, post a photo, make a payment, join a group chat, make a conversation, scan a QR code, etc., all these behaviors change, develop over time, and we call these temporally changing behavioral phenomena of each user as the dynamics of human behaviors, or *Human Dynamics* for short. The WeChat was first released in 2017 with only few users, and then developed to a super huge and complex social network with a billion users and tens of billions of social links. After registering WeChat accounts, new users build social links to the existed users, and recommend the WeChat app to his friends who have not registered yet. We call the dynamic phenomena during the evolving process of the WeChat social network as the *Network Dynamics*. Forming social group is an intrinsic human nature. Over the network, people build all kinds of groups with specific topics. New group members join the group and existed members quit the group, leading to the group evolution dynamics, namely, *Group Dynamics*. Social medias right now are the major platforms for us get informed. The information is produced, spreaded and consumed over the social network, taking on information cascading phenomena. We call the dynamic phenomena of the information cascading process as the *Cascade Dynamics*. We summarize all these dynamic phnomena in the social sytems as *Social Dynamics* (as shown in Fig. 1-Dynamics panel).

Why do the social dynamics matter? The phenomena in WeChat, just like many laws of nature, are of dynamic origins. The studies on the dynamics of WeChat social system, to the best of our knowledge, provide the first-ever and the largest empirical investigations on the real-world dynamics of the human behaviors,

Conference'17, Washington, DC, USA

DOI: 10.1145/nnnnnnn.nnnnnnn

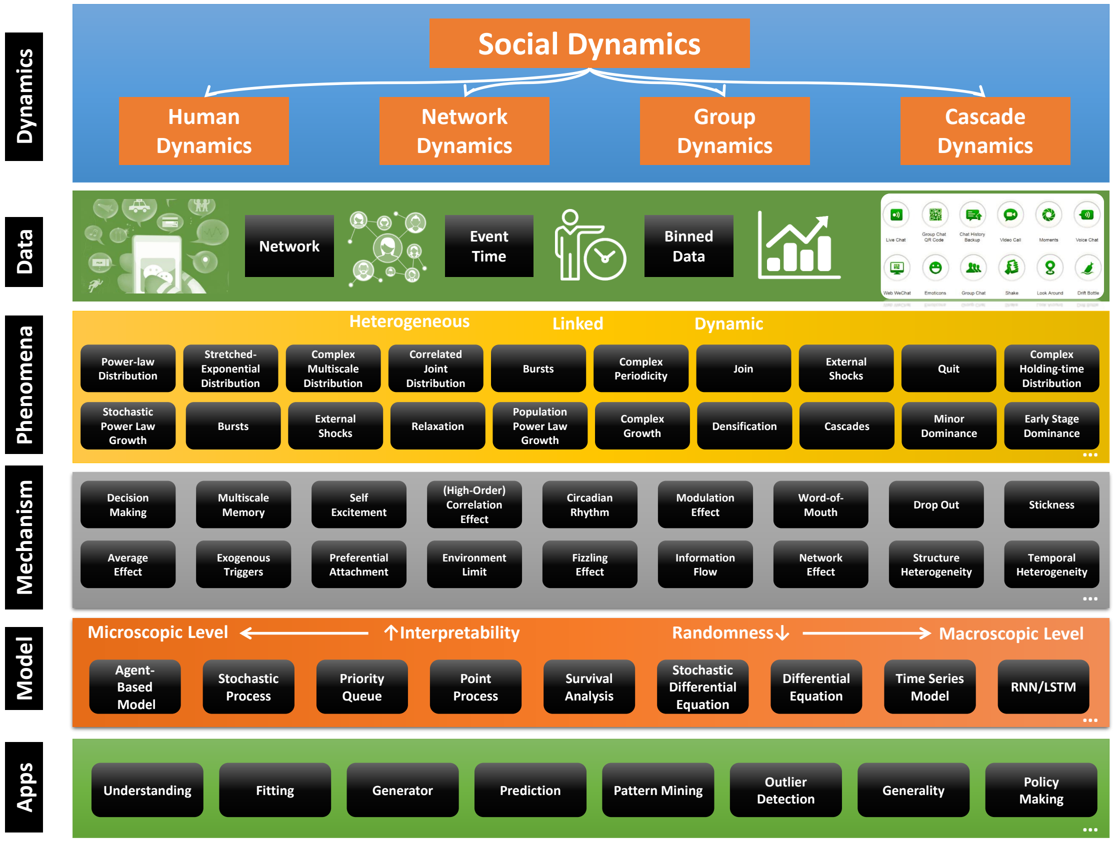

Figure 1: *Research Framework of Modeling Dynamics of WeChat Social System.* The social dynamics are generated by the human behaviors in the nature, including human dynamics, network dynamics, group dynamics and cascade dynamics, etc. The digital records are sampling data of the social dynamics. We attempt to understand and model the social dynamics by understading the phenomena reflected in the data records. Possible mechanisms accouting for the phenomena are the pivot to understand the origin of social dynamics, to choose and design suitable models. The models are also degined and used for specific applications.

network evolution, group evolution and information cascading phenomena. For example, we validate a lot of previous knowledges, ranging from social science, economy, statistical physics to computer science, on the social dynamic phenomena. Some of them are consitent with what we find in WeChat, while many of them disobey the realities we observed. For example, does the spread of WeChat really follow Bass model with S-shaped curve [33]? What's more, indeed, we find a lot of new phenomena in WeChat (such a complex and huge real-world social system), which are not yet known in literature. Understandings of the real-world social systems, including data structure, phenomena, mechanisms, modeling techniques, applications, etc., are very valuable for us to further understand the constantly changing complex social systems in the real world. On the other hand, a lot of issues are of vital importance for the industry. Can we predict the behaviors of next online purchase of each user in WeChat? Can we spot the abonormal groups through their dynamic behaviors, which may possibly recruit terrorist members, or spread illegal informations? How many new users will WeChat (or Facebook, Twitter, etc.) have next month (company valuation)? What are the potential popularity of a new WeChat verion when being released? What is the right timing of releasing next version of WeChat, or an advertisment? Can we predict if an post will be popular or not? Can we apply our knowledge deveolped in WeChat to other social systems? And other usefulness include spreads of social products, provisioning, social implications of policy changes, churn prediction, policy making, and so on. All these questions are social-dynamics-centric.

However, to understand and to model the dynamics of WeChat social system are very challenging. The chanllges lie in the following five folds:

Data Challenge [1] **.** Lack of detailed traces of human dynamics, network dynamics, group dynamics and cascade dynamics prevents previous understanding of social dynamics. The WeChat offers possibely the largest and the most detailed records of various dynamic changes of a real-world social systems. As shown in the Fig. 1-Data panel, we have the network data, the time of each event, and the aggregated data in various social scenarios. However, challenges coexist with opportunities. The records generated in WeChat, including the network data, behavior data and so on, are indeed "big data". Besides, these "big data" are indeed complex, ranging from structure data to temporal data, from individual data at microscopic level to population data at macroscopic level, from network evolution data, chatting behavior data, group formation data to information flow data, covering different social scenarios.

Phenomena Challenge. Refering to the main framework illustrated in Fig. 1, the social dynamics are generated by the human

[1] All the data that we could access were behavior data with anonymized identities. No content data can be reached. Strict privacy policies are followed.

behaviors in the nature, the digital records are sampling data of the social dynamics, and we attempt to understand and model the social dynamics by understading the phenomena reflected in the digital records. However, we observe quite complex phenomena, as enumerated in the Fig. 1-Phenomena panel, ranging from structure phenomena to temporal phenomena, from individual phenomena at microscopic level to population phenomena at macroscopic level, from individual behavior phenomena, network evolution phenomena, group evolution phenomena to information flow phenomena, covering different social scenarios. The complex phenomena are heterogeneous, linked and dynamic.

Mechanism Challenge. Underlying the complex phenomena are the complex mechanisms which governing human and social behaviors, as shown in the Fig. 1-Mechanism panel. We cannot emphasize enough the importance of mechanism understanding. First, to understand the mechanisms is the major road to uncover social dynamics and social systems. Second, to understand mechanisms is the vital pivot for us to choose and design models to capture, predict and change the realities. However, to understand intrinsic mechanisms underlying complex phenomena are very challenging. We are limited and biased by what we learn, what we see, what tools we use, and how we interpret.

Model Challenge. Faced with such complex data, phenomena, mechanisms and applications (discussed later), how to choose suitable model framework and then design specific models are very challenging. Is there a model which can capture and reproduce the phenoemena ranging from structure dimension to temporal dimension, from individual at microscopic level to population at macroscopic level, from individual behavior, network evolution, group evolution to information flow, covering different social scenarios, and be applied to different applications? Possibly the ansewer is NO.

Application Challenge. Actually in the "Why do the social dynamics matter" paragraph, we illustrate a lot of questions or applications the academia and the industry may concern about. These applications are proposed by different people and different teams who care about different social scenarios, ranging from human dynamics, network dynamics, group dynamics and cascades dynamics, from microscopic level to macroscopic level. How to cluster and abstract each specific task into a higher level problem is of vital importance and challenge. Besides, different models are usually used for different tasks. How to choose suitable model framework for specific applications are challenging.

Our Contributions. For the first time, we try to summarize our experiences in dealing with social dynamic phenomena, data and related problems occurring in (large-scale, real-world, complex) social systems, especially in WeChat social systems, into a principled framework (Fig. 1). Four major aspects of social dynamics are discussed (Figs. 2-4), including human dynamics, network dynamics, group dynamics, and cascade dynamics. We abstract the data records in these four social dynamic scenarios into data structures. After that, we summarize the new phenomena we observed in WeChat data and other social systems. We explain our understanding of each phenomenon, and investigate the complex relationships underlying these phenomena in different social dynamic scenarios.

We try our best to leverage interdisciplinary knowledges, covering social science, statistical physics, economy and computer science, to understand and interpret the possible mechanisms underlying the phenomena we observed and cared about. We explain what these mechanisms are. Furthermore, we show the complex "network" underlying our research framework in the four social dynamic scenarios, including the relationships within these mechanisms, the relationships between phenomena and mechanisms, and the models which are suitable to capture the targeted phenomna based fundamentally on these mechanisms (Figs. 2-4).

We give a spectrum of modeling frameworks (methodology) which can be potentially used to capture dynamic phenomena in different realms, as shown in the Fig. 1-Model panel, ranging from agent-based model (used in physics and social science), stochastic process (physics and math), queuing theory (math and engineer), point process (math), survival analysis (statistics, epidemiology), stochastic differential equation (finance), differential equation (physics and engineer), timeseries model (engineer and computer science), deep learning like RNN and LSTM (computer science) and other techniques developed in data mining and machine learning fields. We summarize this spetrum of modeling framwroks according to their abilities of capturing microscopic phenomena or macroscopic phenomena, stochastic phenomena or deterministic phenomena, temporal phenomena or structure phenomena, and their interpretability. Usually, as shown in the spetrum of methodologies in the Fig. 1-Model panel, if we want to capture the microscopic and stochastic phenomena, possibly with network informations and with good interpretability, we tend to seach models leftward. On the other hand, if we want to capture the macroscopic and population level phenomena, with less randomness, less interpretability, and less emphasis on the intrinsic complexity and heterogeneity among the population, we tend to models rightward. We show our choice of the methodology and model design in each following section.

We summarize the commonly asked questions and applications as: understanding, fitting, generator (simulation), prediction, pattern mining, outlier detection, generality, policy making and so on. The research framework we give, illustrated in Figs. 1-4, could potentially guide future researches on social dynamics related problems, sheding light on modeling and designing data structure, explaining observed phenomena, inferring possible mechanisms, choosing and design dynamic models and pinpointing applications in the real-world social systems.

2 HUMAN DYNAMICS

The human dynamics try to capture the dynamics of individual human behaviors, focusing on the timing of actions and their intrinsic driven factors. For example, when will you add a new friend? When will you response to your frineds in a conversation or a group chat? When will you post an orginal tweet? When will you retweet a message from your friends? When will you start or response an email? When will you purchase another prop in an online game? By following the pattern, i.e. when will sb do something, we can enumerate hundreds of "WHEN" questions in the WeChat ecosystem. Furthermore, we want to know "WHY" these behaviors occur, namely, mechanisms. Thus, the studies on human dynamics are the building block of all kinds of social dynamic phenomena, and we try

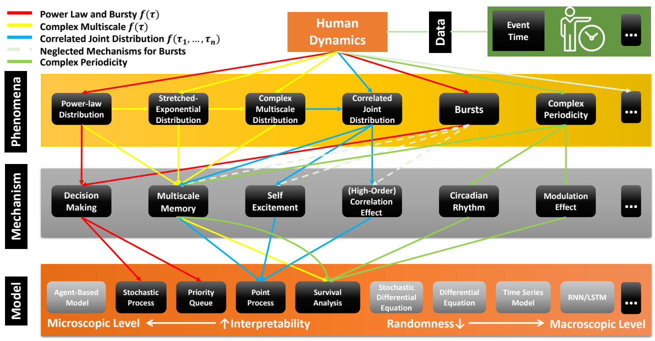

Figure 2: *Human Dynamics.* Our research network underlying the human dynamics. Phenomena and corresponding mechanisms and models are linked by arrow lines with same color. The links in the phenomena panel illustrate the progress of researches based on new phenomena and thus new understandings.

to find and model the shared phenomena among above questions. We illustrate the "research network" of human dynamics in Fig. 2.

2.1 Data

Usually, the data of the human dynamics are recorded as a sequence of event time $\mathcal{H}_n = (t_1, t_2, \ldots, t_{n-1}, t_n)$, where t_i is the time of the i^{th} event. Some equivalent transformations of $\mathcal{H}_n$ are:

- Couting process $N(t) := \sum_{t_i \leq t} 1_{(0,t]}(t_i)$, or a sequence $(t_1 : 1, t_2 : 2, \ldots, t_n : n)$.
- Inter-event time (IET) sequence $\mathcal{T}_n(\tau_1, \tau_2, \ldots, \tau_{n-1})$, where $\tau_i = t_{i+1} - t_i, i = 1, \ldots, n-1$.

2.2 Phenomena

Due to the intrinsic complexity of human behaviors, the understanding and modeling of human dynamics starts from investigating the statistical properties of inter-event time, i.e., the probability denstity function $f(\tau)$. The seminar studies on human dynamics found that the *i) inter-event time* τ follows *ii) power law distribution* $f(\tau) \propto \tau^{-\alpha}$, leading to *iii) bursty behaviors* [5, 27]. However, the reality exhibits much more complexites:

(1) *Complex multiscale distributions.* We find the inter-event time distribution $f(\tau)$ exhibits complex multiscale distribution rather than the pure power law distribution in various datasets [32, 36]. Besides, the power-law distribution is indistinguishable from stretched-exponential (Weibull) distribution, or other heavy-tailed distributions [9].
(2) *Correlated joint distribution.* Furthermore, we found that the inter-event times are correlated, i.e., $f(\tau_i, \tau_{i+1}) \neq f(\tau_i)f(\tau_{i+1})$ or even high-order-temporal correlations , i.e., $f(\tau_i, \tau_{i+k}) \neq f(\tau_i)f(\tau_{i+k})$, beyond consecutive inter-event times [32], indicating that modeling of $f(\tau)$ only is inadequate [5, 27]. Thus we try to capture the $\mathcal{H}_n$ ($N(t)$ or $\mathcal{T}_n$) directly rather then $f(\tau)$ only.
(3) *Complex periodicity.* We observe complex periodical changes of $\mathcal{T}_n$ [29]. For example. the dynamics of we sending or reading messages change periodically during a day or a week.

2.3 Mechanisms and Models

The human dynamics describe the individual behaviors at microscopic level, exhibiting large randomness. Besides, the modeling of human dynamics possibly serves as the microscopic building block of other dynamics, and we'd better give it a good interpretation. Thus, we prefer model frameworks towards the microscopic direction as shown in Fig. 2.

(1) *Complex multiscale distributions.* The physicists explain the power law $f(\tau)$ by the decision making process, and they model this process by priority queue from queueing thoery or stochastic process [27]. As for the complex multiscale distribution, however, there is no such a model in literature. We try to make it a user-friendly building block for the following models in more complex scenarios, and we choose the tools developed in the survival analysis [16] to model the complex distribution [36]. By modeling the hazard rate $\lambda(\tau)$ which captures the multiscale memory of human behaviors [36], we get complex distributions according to the formula $f(\tau) = \lambda(\tau)e^{-\int_{-\infty}^{t} \lambda(s)ds}$, together with principled ways to learn the parameters and simulation random samples developed in the hazard rate space.
(2) *Correlated joint distribution.* The correlated joint distribution of inter-event times are due to the high-order-temporal correlations, or self excitement of previous behaviors [32]. However, the survival analysis models are designed for the *i.i.d.* random variables. Thus, we resort to more general tools which can model the high-order-temporal correlations and at the same time encompass the results of complex multiscale distributions modeled by hazard rate. We choose point process, say Hawkes process [17] and its variants [32], to model the correlated events. Detailed in the network dynamics section.
(3) *Complex periodicity.* The complex periodicity exhibits in different time scale, from second, minute, hour, day, week to month, etc. These complex periodicity are derived from the multiscale memory modulated by the circadian rhythm of human behaviors [29]. We follow the point process framework and model the periodicity by a mixture model with 7 components to capture weekly periodicity, each one of which captures the daily periodicity [29]. The modeling of weekly periodicity and daily periodicity strikes a balance between the simplicity of models and the accuracy of modeling complex periodical dynamics at a reasonable time scale.
(4) *Bursty* $f(\tau)$. Previous explaination of bursty phenomena is that the burst is derived from heavy-tailed $f(\tau)$ only [5]. However, other mechanisms, including self-excitement, (high-order) temporal correlation also lead to bursts. We discuss this in the network dynamics section.

3 NETWORK DYNAMICS

The network dynamics are the dynamics of the evolving process of social networks, at both individual level and population level. As the largest social network in China, to study the network dynamics of WeChat is of vital importance for the academic interests: How does the real-world social network look like? How does WeChat

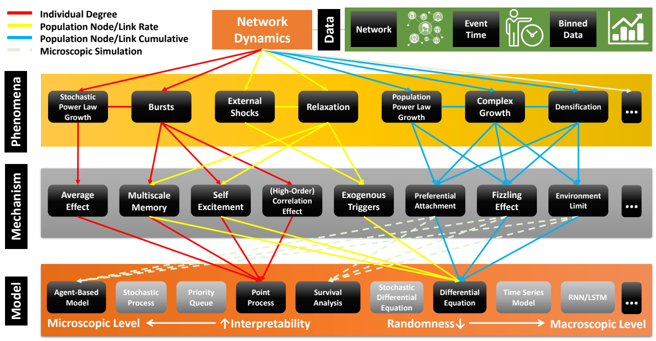

Figure 3: *Network Dynamics.* Our research network underlying the network dynamics. Phenomena and corresponding mechanisms and models are linked by arrow lines with same color.

grow over time? Do the network dynamics of WeChat follow or disobey previous assumptions? Besides, the studies on network dynamics of WeChat are also very import to industrial concerns: How many new users will WeChat (or Facebook, Twitter, etc.) have next month? How many social links are there next year? How does each user add friends over time? How many new users will use WeChat when Tencent releases a latest WeChat verion? etc. All these question are crucial for the company valuation, spreads of social products, provisioning, social implications of policy changes, churn prediction etc.

We marjorly care about following four questions on network dynamics of WeChat:

(1) How does the number of friends $k_i(t)$ of user i change over time?
(2) How does the number of new users $n(t)$ change over time after the release of a new WeChat version at t_0?
(3) How does the totoal number of users $N(t)$ in WeChat grow over time?
(4) How does the total number of social links $E(t)$ in WeChat grow over time?

We see the problem (1) is at microscopic level, namely for each individual, while problem (2-4) are at macroscopic level, namely for population. The intrinsic differences between the individual dynamics and populations dynamics lie in the differences in data forms, phenomena, mechanisms and models, as shown in Fig. 3

3.1 Data

For each user i, we know the records $(i, j, t_{i,j}), j = 1, \ldots, k_i$ which represent the user i add user j as his/her friend at time $t_{i,j}$. Logically, the social link in WeChat is bidirectional. Thus, we define the history of the adding friend dynamics of user i as $\mathcal{H}(i, k) = (t_1, \ldots, t_{k-1}, t_k)$.

- The individual degree dynamics, namely the number of friends of user i, is a counting process $k_i(t) := \sum_{j \geq 1} 1_{(0,t]}(t_j)$.

On the other hand, we also care about the population dynamics, recorded as the binned number over time $t = 1, \ldots, T$ with time scale hour, day, week or quarter etc., including:

- $n(t)$: the number of new WeChat user at time t;
- $N(t)$: the cumulative WeChat user by time t, where $N(t) = \sum n(t)$;
- $e(t)$: the number of new social links in WeChat at time t;
- $E(t)$: the cumulative social links in WeChat user by time t, where $E(t) = \sum e(t)$;.

3.2 Phenomena

The network dynamics exhibit large complexities at different level, ranging from individual level to population level.

(1) *Individual degree dynamics $k_i(t)$.*
   - *Stochastic power law growth.* We find $k_i(t)$ exhibits various nonlinear growth patterns at long term, including accelerating power-law growth, linear growth, decelerating power-law growth, etc. [32]. For example, active users or even outlier users exhibit accelerating growth of $k_i(t)$. Besides, the dynamics of individuals exhibit large randomness.
   - *Bursts.* The $k_i(t)$ exhibits bursty growth (bursts) at short term[32]. For example, in China, we exchange WeChat accounts instead of exchanging printed name cards. When we attend a conference, we may add a lot of new friends within a short period of time, leading to bursts in $k_i(t)$. However, the bursts are derived from complex mechanisms rather than $f(\tau)$ only.

(2) *Population rate $n(t)$* . We find $n(t)$ exhibits recurring patterns of external shocks followed by a long period of relaxation dynamics [under review]:
   - *External shocks.* We find a lot of dynamic changes are triggered by external shocks. For example, the release of a major version of WeChat may inccur large shocks of $n(t)$. The promotion acitivities like the digital red packet during Chinese New Year also leads to a large shock of $n(t)$. These phenomena are very common, because WeChat is updated for iPhone, Android, Windows Phone, etc., almost every month.
   - *Relaxation.* We also find after the external shock is a long period of relaxation regime, which accounts for the majority of the lifetime between consecutive two shocks.

(3) *Population cumulative number $N(t)$ and $E(t)$.* By examing the cumulative number of nodes $N(t)$ over time, we focus on trend rather than randomness.
   - *(Population) Power law growth.* The previous network model [6, 19] all assume uniform growth of nodes $N(t)$. And, the celebrated Bass model [4, 20] which is widely used to model the spreads of innovation, and the Susceptible-Infected (SI) model [2] which is used to model the epidemic spreading, all produce Sigmoid (S-) curve with exponential early growth of $N(t)$. And none of them studied the dynamics of link $E(t)$. However, reality disagrees. We find power law early growth of both node $N(t)$ and link $E(t)$, indicating their scaling relationships over time, implying a rather long period of consistent growth [31, 33].
   - *Complex growth pattern.* However, the power law growth may reach the inflection point due to limited potential population. And the scaling relationship is origined from the critical state which is seldom

reached strictly. Thus, the real growth pattern are complex, possibly taking on stretched long-term growth pattern before reach the upper limit [31, 33].

- *Densification of Links.* The growth of links are largely neglected previously. The links built between the already-existing nodes leading to densification phenomenon. Further, we find the $E(t)$ scales with $N(t)$ in WeChat, taking on $E(t) \propto N(t)^{1.41}$ [31, 33]. .

3.3 Mechanisms and Models

Here we try to figure out what are the mechanisms governing the phenomena we observed, and what are the suitable model to capture these mechanisms. The thingkings leading to the model are shown.

(1) *Individual degree dynamics $k_i(t)$.*

- *Average effect and Stochastic power law growth.*
  - *Average effect.* Starting from the macroscopic level model, we find the power law growth is due to the *average effect*, which describes the rate of making friends $\frac{dk_i(t)}{dt}$ at present is proportional to the long-term average rate $\frac{k_i(t)}{t}$, and this long-term memory is kept by the behavior of making $k_i(t)$ friends so far, namely: $\frac{dk_i(t)}{dt} = \frac{\alpha}{t} k_i(t)$. By solving above equation, we can get $k_i(t) = k_i(t_0)(\frac{t}{t_0})^{\alpha}$. However, this is a deterministic equation which can not capture the intrinsic randomness in the individual behaviors.
  - *Stochastic model.* By searching the stochastic model towards the microscopic level direction as shown in Fig. 3, we choose the point process (more specifically the counting process) to model $k_i(t)$. The reasons are in three folds: *a)* the stochastic differential equation models usually capture the randomness derived from population level rather than individual level; *b)* the survival analysis model, as discussed above, needs to be generalized its i.i.d. assumptions to the non- i.i.d. case; *c)* the number of friends $k_i(t)$ is a stochastic couting process over time. Thus, we give *a (stochastic) power-law process*, modeled by intensity $\lambda(t|\mathcal{H}_t) = \lambda_\infty \alpha(\frac{t}{\Delta_\infty} + 1)^{\alpha-1}$ [32].
- *Multiscale memory, Self-excitement, High-order temporal correlation and Bursts.* The bursts are complex phenomena which may be orgined from multiple mechanims which serve as the building blocks of the model:
  - *Multiscale memory.* In the human dynamics section, we show the human behaviors often exhibit multiscale inter-event-time (IET) distribution [36] taking on flat short-term-memory, heavy-tailed middle-term memory, and the exponential-like long-term memory since last event. When we i.i.d. sample IETs from a multiscale distribution, a sequence of small value IETs occur in short-scale, followed by large value IET in the middle-scale and long-scale, generating intense activities followed by a long vacation, namely, bursts. However, due to ignoring the correlations between IETs, this is an i.i.d. $f(\tau)$ bursty mechanism.
  - *(High-order-temporal) Correlation effect and Self-Excitement.* Given a Poisson process exhibiting non-bursty growth, if we rearrange the IET sequence as follows: make small value IET followed by smaller IET, and make large value IET followed by larger IET, and then we also get bursty behaviors. Thus, the correlations between IETs can also generate bursts. Studies on Hawkes process shed light on one possible explanation of these (high-order-temporal) correlation effect : the self excitement, describing cumulation of the endogenous self-excitement of previous events. The intensity function of the Hawkes model is $\lambda(t|\mathcal{H}_t) = \mu + \sum_{t_i<t} \alpha(t - t_i)^{-\beta}$ [17].

Based on the above discussion, we give the Long-Short-Term Memory Process (LSTMP) to capture $k_i(t)$, which generalizes the assumptions of the Hawkes process, namely, long-term perturbation term μ to power-law process, the infinite memory length to tunable memory length, and power-law kernel to complex multiscale kernel, leading to: The hazard function specifying the LSTMP is: $\lambda(t|\mathcal{H}_t) = \lambda_\infty \alpha(\frac{t}{\Delta_\infty} + 1)^{\alpha-1} + \sum_{i=n(t)-m+1}^{n(t)} \lambda_0(\frac{t-t_i}{\Delta_0} + 1)^{-\theta}$ [32].

(2) *Population rate $n(t)$.* Now we turn to explain how the number of new users $n(t)$ changes over time after the release of new WeChat version at t_0. External shocks are triggered by the exogenous events. Then the WeChat social system responses to it, taking on relaxation/decay of $n(t)$ since $n(t_0)$. Inspired by the individual dynamics $k_i(t)$, we model the social system as a Hawkes process like system (thus with multiscale memory, self-excitement mechanisms). However, we need to adapt it to model population level dynamics $n(t)$, and incorporate external shocks. Thus, we take mean-field average of the intensity function $\lambda(t|\mathcal{H}_t) = \mu(t) + \sum_{i,t_i<t} \beta_i \frac{1}{(t-t_i+\Delta)^{\theta}}$ to $n(t) = n_0(t) + \beta \int_{t_0}^{t} \frac{1}{(t-\tau+\Delta)^{\theta}} d\tau n(\tau)$, and model the relaxation dynamics $n(t)$ after t_0 by differential equation models $n(t) = n_0(t) - \beta \int_{t_0}^{t} \frac{1}{(\tau-t_0+\Delta)^{\theta}} d\tau n(\tau)$, and its variants [under review].

(3) *Population cumulative number $N(t)$ and $E(t)$.* The $N(t)$ and $E(t)$ are governed by three mechanisms: preferential attachment, environment limit, and fizzling effect [31, 33].

- *Preferential attachment.* A social network with a large population $N(t)$ has a propensity to attract more nodes in the early stage. The microscopic explanation is like an epidemic model: existed/infected users ask/infect new/susceptible users to adopt WeChat.
- *Environment limit.* As the population who can join the social network is limited, its growth will be constrained by the decreasing number of potential nodes $(N - N(t))$, especially at the saturation stage. Total

smart-phone users possibly serve as the total potential population N to be infected. As for the links, total possible links one user can build are also limited, due to cognition, social and economic reasons.

- *Fizzling effect.* The infection rate fizzles over time, possibly due to the competition of other social products, fogetting nature of human interets, etc., taking on fizzling effect.

In all, we model the totoal number of nodes $N(t)$ and links $E(t)$ by differential equations $\frac{dN(t)}{dt} = \frac{\beta}{t^{\theta}}N(t)(N - N(t))$ and $\frac{dE(t)}{dt} = \frac{\beta'}{t^{\theta}}N(t)(\alpha(N(t)-1)^{\gamma} - \frac{E(t)}{N(t)}) + 2\frac{dN(t)}{dt}$ respectively to capture the network dynamics at *population level.* Our node model uncovers various growth framework, including exponential growth, power law growtrh, stretched-exponential growth, Sigmoid (Logistic) growth, log-logistic growth and stretched-logistic growth. Our link model captures the power law growth of links and densification power law. We further explain our population level model by agent-based simulation and survival analysis model [31, 33].

4 GROUP DYNAMICS

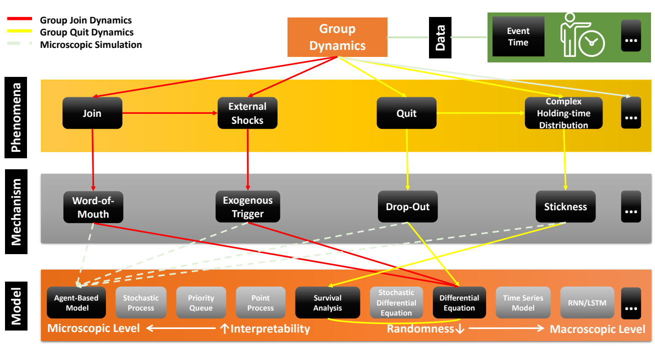

Figure 4: *Group Dynamics.* Our research network underlying the group dynamics. Phenomena and corresponding mechanisms and models are linked by arrow lines with same color.

The group dynamics are the evolution dynamics of social groups, which are the aggregated dynamics of individual behaviors which are associated with a specific group (say, family group, classmates group, bussiness group, work team group, interest group, alumni group, etc.) in WeChat. The group members join and quit groups, leading to group evolution dynamics. For example, we want to know how many members will this bussiness group have next week? How many members will quit the group next month? The study of group dynamics is of vital importance. First, forming social group is an intrinsic human nature. Second, the group dynamics may influence human dynamics, network dynamics, and information flow dynamics. Third, groups have semantics. The study of business group dynamics can provide insight into the evolution of organizational structure of economic behaviors; the study of abnormal group dynamics can detect outlier behaviors, say the online terrorist recruitment, and so on. We illustrate the "research network" of group dynamics in Fig. 4.

4.1 Data

For each group g, we have the join records $\mathcal{J}_{g,n} = \{(t_1, u_{J_1}), ..., (t_n, u_{J_n})\}$, where t_i is the time of the i^{th} user u_{J_i} who joins the group g, and the quit records $\mathcal{Q}_{g,m} = \{(t_1, u_{Q_1}), ..., (t_m, u_{Q_m})\}$, where t_j is the time of the j^{th} user u_{Q_j} who quits the group g. Similarly, we have two counting process which desscribe the join process and quit process respectively:

- Holding time τ_i of user i, the time interval between i quit and join time, $\tau_i = t_{Q_i} - t_{J_i}$.
- Join couting process $J(t) := \sum_{t_i \le t} 1_{(0,t]}(t_i)$, or a sequence $(t_1 : 1, t_2 : 2, ..., t_n : n)$.
- Quit couting process $Q(t) := \sum_{t_j \le t} 1_{(0,t]}(t_j)$, or a sequence $(t_1 : 1, t_2 : 2, ..., t_m : m)$.

4.2 Phenomena

According to the data we have, we can decompose the group dynamics into join group dynamics and quit group dynamics. The group dynamics capture the mesoscopic dynamics over the network, implying mixture characteristics of both microscopic dynamics and macroscopic dynamics. Specifically, we find the join group dynamics exhibit [35]:

(1) *Diffusion join dynamics.* The existed group members invite outsiders to join the group, exhibiting diffusion-like (word-of-mouth, or epidemic-like) join dynamics.
(2) *External-shock join dynamics.* Besides, a bunch of outsiders may join the group simultaneously due to some external triggers, taking on shock-like dynamics.
(3) *Complex quit dynamics.* How members quit the group is largely unknown. We find that the holding time (the time between join and quit the group) distributions for different group exhibit large complexity and heterogeneity, ranging from exponential-like distribution, power-law-like distribution and inbetween cases.

4.3 Mechanisms and Models

We find the join process exhibits collective dynamics while quit process exhibits large randomness. Usually, a person joins a group due to invitations, while quits a group largely by the decision of himself. Thus, we choose the differential equation, survival analysis to capture the join and quit group dynamics, and agent-based model to simulate microscopic process as shown in Fig. 4.

(1) *Diffusion join dynamics.* The diffusion dynamics are due to word-of-mouth, or epidemic-like process. We model the collective dynamics by differential equation inspired by Susceptible-Infected (SI) model and Susceptible-Infected-Recovered (SIR) model. As for the microscopic simulation, we follow the agent-based model of SI and SIR [2].
(2) *External-shock join dynamics.* In WeChat, we have a lot of ways to join a group. A bunch of outsiders may join the group simultaneously due to some external triggers, like joining a new built group in a conference by face-to-face function, or scaning an online QR-code to join the group,

etc. We model these shocks by adding mixture of dirac delta functions into the differential equation model.

(3) *Complex quit dynamics.* The most complex and interesting phenomena are the quit dynamics. New members join group collectively, however, existed members quit the group largely due to their own decisions, leading to large randomness. Furthermore, we find the stickness of staying in a group, indicating an existed member tends to stay in the group longer if he/she has already stayed there for a long time. In contrast, new members tend to quit the group quickly. We model this temporal stickness by survival analysis model and then incorporate it to the differential equation model [35].

The combination of population dynamics and individual dynamics are the key characteristics of group dynamics. Thus, for the first time, we combine the differential equation models and survival analysis models to capture mesoscopic group dynamics[35].

5 CASCADE DYNAMICS

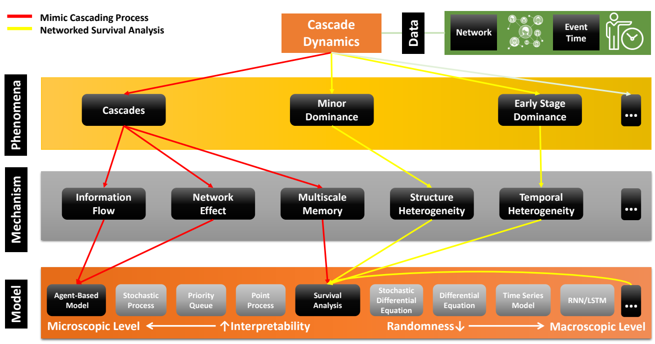

Figure 5: *Cascade Dynamics.* Our research network underlying the information cascade dynamics. Phenomena and corresponding mechanisms and models are linked by arrow lines with same color.

The cascade dynamics are the dynamic cascading process of information over networks. For example, in WeChat or any other social networks, the information you received is from one of your friend, and you want to share it to your other friends, taking on information cascading phenomena, or cascades for short. Cascades are very important in various network scenarios, partly because of their ubiquity, and partly because of their various applications, ranging from viral marketing, epidemic prevention, to predicting the spreads of information. In WeChat and other Tencent social systems, say Tencent Weibo, we focus on how to predict the cascading process, That is, given the early stage of an information cascade, can we predict its cumulative cascade size at any time [30]?

5.1 Data

Given a network $G = \langle \mathcal{N}, \mathcal{E} \rangle$, where $\mathcal{N}$ is a set of nodes and $\mathcal{E}$ is a set of directed/undirected relationships. A piece of information can be originated from one node and spread to its neighboring nodes. A cascade is typically formed by repeating this process. Therefore, a cascade can be represented as following data structure: $C = \{(u_1, Par(u_1), t_1), (u_2, Par(u_2), t_2), ..., (u_m, Par(u_m), t_m)\}$, where u_1 is the root node, $Par(u_i)$ is the user who passed the information to u_i, denoted as the "parent" user, and $t_1 \le t_2 \le ... \le t_m$. The cascade data structure contains both structure and dynamic informations [34]:

- Cascade structure: $C_s = \{(u_1, Par(u_1)), ..., (u_m, Par(u_m)\}$.
- Cascade dynamics: $C_d = \{t_1, t_2, ..., t_m\}$, with counting process $C_d(t) := \sum_{t_j \le t_m} 1_{(0,t]}(t_j)$, or a sequence $(t_1 : 1, t_2 : 2, ..., t_m : m)$.

5.2 Phenomena

The dynamic spreading process of information flow over network is a complex phenomenon:

(1) *Cascading Phenomenon.* The information a user may get is influenced by his/her friends, and thus cascading phenomenon occurs naturally.

(2) *Minor dominance.* Although a information cascades reaches a lot of people, each one in the cascades contributes to the total cascade size significantly different. It is intuitive that an active user with thousands of friends can influence more friends than that of an inactive user with several friends. The size of a cascade is usually dominated by few nodes, taking on star-like structure.

(3) *Early-stage dominance.* We further find the dominant nodes are also prone to join cascades in early stage, taking on bursty growth dynamics.

5.3 Mechanisms and Models

Predictive modeling on cascades has aroused considerable interests recently. Earlier works focus on predicting the final size based on feature engineering on content, behavior dynamics and structure [8], important nodes [11], and so on. More ambitiously, we attempt to predict the evolving process of a cascade. A fundamental way to address this problem is to look into the microscopic mechanisms of how cascades take place.

(1) *Decomposing cascading phenomenon.* In order to capture the cascading phenomenon, we decompose the macro- cascade into multiple one-hop subcascades. We adopt the survival model to capture dynamics of each subcascade, and then sum up each subcascades over the real network by an agent-based model.

(2) *Dominance and heterogeneity.* The minor dominace and early-stage dominance are derived from structure heterogeneity and temporal heterogeneity respectively. In order to capture the temporal heterogeneity, we use Weibull hazard rate to capture the heterogeneous temporal dynamics with different roo users. However, the traditional hazard model suffers two drawbacks in cascade scenario: *a)* some users have very sparse or even no data on their subcascades; (2) the hazard rates of connected nodes are also correlated. Thus, we can not directly apply the survival model to the cascading phenomenon. To address these, we leverage the knowledge in social science and data mining methods, introducing interpretable covariates of each user, say their structure, behavior, etc., features, to regularize the hazard rate. We then propose a Networked Weibull Regression model to capture above ideas [30].

We combine network-based simulation, survival analysis with regularizers to capture cascading process.

6 APPLICATIONS

The fitting, generator/simulation, and prediction applications are the three basic methods to validate the proposed models in all the scenarios. However, by following the general methods like time series models at macroscopic level, it's difficult for us to interpret, understand and to do policy making by human decisions. Besides, they are not suitable to model non-linear, stochastic, and bursty social dynamics. However, by following our frameworks in all four social dynamic scenarios, pivoted by phenomena and mechanism procedures, our models give quite good performance, and at the same time provide interpretations for various scenarios we considered. Besides, pattern mining and outlier detection are conducted in the parameter space with interpratable physical meanings [30–33, 35, 36]. As for the generality, our dynamic models are applied to a wide range of social systems [30–33, 35, 36].

7 CONCLUSIONS

In this paper, we systematically summarize our experiences in dealing with social dynamic phenomena, data, and related problems in WeChat social system, which is one of the largest, most complex social systems in the real world. A framework covering the pipeline of dealing with complex data, complex phenomena, complex mechanisms, a spectrum of methodologies, and various applications is proposed. By following our framework, we mainly discuss four aspects of social dynamics, including human dynamics, network dynamics, group dynamics and cascade dynamics. We abstract the data structures from the digital logs, investigate the phenomena we observed from the data, explain the possible mechanisms governing the phenomena, and give the models we choose and design according to our methodology spectrum of modeling dynamics. We illustrate the specific "research network" of our reseach frameworks when applied to four dynamic scenarios. Typical applications are summarized and illustrated. The framework could potentially guide future researches on social dynamics related problems, sheding light on modeling and designing data structures, explaining observed phenomena, inferring possible mechanisms, choosing and designing dynamic models, and pinpointing applications in the real-world social systems.

REFERENCES

[1] Rodrigo Augusto da Silva Alves, Renato Martins Assuncao, and Pedro Olmo Stancioli Vaz de Melo. 2016. Burstiness Scale: A Parsimonious Model for Characterizing Random Series of Events. In *Proceedings of the 22nd ACM SIGKDD*. ACM, 1405–1414.

[2] Roy M Anderson, Robert M May, and B Anderson. 1992. *Infectious diseases of humans: dynamics and control*. Vol. 28. Wiley Online Library.

[3] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin Dunbar. 2013. Dynamics of personal social relationships in online social networks: a study on twitter. In *Proceedings of the first ACM conference on Online social networks*. ACM, 15–26.

[4] Robert B Banks. 1994. *Growth and diffusion phenomena: mathematical frameworks and applications*. Vol. 14. Springer Science & Business Media.

[5] Albert-Laszlo Barabasi. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435, 7039 (2005), 207–211.

[6] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

[7] Ginestra Bianconi and A-L Barabási. 2001. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)* 54, 4 (2001), 436.

[8] 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.

[9] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review* 51, 4 (2009), 661–703.

[10] Riley Crane and Didier Sornette. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *PNAS* 105, 41 (2008), 15649–15653.

[11] Peng Cui, Shifei Jin, Linyun Yu, Fei Wang, Wenwu Zhu, and Shiqiang Yang. 2013. Cascading outbreak prediction in networks: a data-driven approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 901–909.

[12] Angelos Dassios, Hongbiao Zhao, and others. 2013. Exact simulation of Hawkes process with exponentially decaying intensity. *Electronic Communications in Probability* 18, 62 (2013), 1–13.

[13] Alceu Ferraz Costa, Yuto Yamaguchi, Agma Juci Machado Traina, Caetano Traina Jr, and Christos Faloutsos. 2015. Rsc: Mining and modeling temporal activity in social media. In *SIGKDD'15*. ACM.

[14] Alan G Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58, 1 (1971), 83–90.

[15] Hang-Hyun Jo, Juan I Perotti, Kimmo Kaski, and János Kertész. 2015. Correlated bursts and the role of memory range. *Physical Review E* (2015).

[16] David G Kleinbaum and Mitchel Klein. 2010. *Survival analysis*. Vol. 3. Springer.

[17] Patrick J Laub, Thomas Taimre, and Philip K Pollett. 2015. Hawkes processes. *arXiv preprint arXiv:1507.02822* (2015).

[18] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. 2008. Microscopic evolution of social networks *(KDD '08)*. ACM, 462–470.

[19] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2007. Graph evolution: Densification and shrinking diameters. *TKDD* (2007).

[20] Vijay Mahajan, Eitan Muller, and Frank M Bass. 1990. New product diffusion models in marketing: A review and directions for research. *The journal of marketing* (1990), 1–26.

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

[22] Jorge J Moré. 1978. The Levenberg-Marquardt algorithm: implementation and theory. In *Numerical analysis*. Springer, 105–116.

[23] Vincenzo Nicosia, Petra E Vértes, William R Schafer, Vito Latora, and Edward T Bullmore. 2013. Phase transition in the economically modeled growth of a cellular nervous system. *PNAS* 110, 19 (2013), 7880–7885.

[24] Tohru Ozaki. 1979. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics* 31, 1 (1979), 145–155.

[25] JA Peacock. 1983. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society* 202, 3 (1983), 615–627.

[26] Pedro Olmo S Vaz de Melo, Christos Faloutsos, Renato Assunção, and Antonio Loureiro. 2013. The self-feeding process: a unifying model for communication dynamics in the web. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1319–1330.

[27] Alexei Vázquez, Joao Gama Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor, and Albert-László Barabási. 2006. Modeling bursts and heavy tails in human dynamics. *Physical Review E* 73, 3 (2006), 036127.

[28] Ye Wu, Changsong Zhou, Jinghua Xiao, Jürgen Kurths, and Hans Joachim Schellnhuber. 2010. Evidence for a bimodal distribution in human communication. *PNAS* 107, 44 (2010), 18803–18808.

[29] Linyun Yu, Peng Cui, Chaoming Song, Tianyang Zhang, and Shiqiang Yang. 2017. A Temporally Heterogeneous Survival Framework with Application to Social Behavior Dynamics. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1295–1304.

[30] 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 *IEEE International Conference on Data Mining, ICDM*.

[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 (KDD '16)*. ACM, 2015–2024.

[32] Chengxi Zang, Peng Cui, Christos Faloutsos, and Wenwu Zhu. 2017. Long Short Memory Process: Modeling Growth Dynamics of Microscopic Social Connectivity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 565–574.

[33] C. Zang, P. Cui, C. Faloutsos, and W. Zhu. 2018. On Power Law Growth of Social Networks. *IEEE Transactions on Knowledge and Data Engineering* PP, 99 (2018), 1–1. DOI: http://dx.doi.org/10.1109/TKDE.2018.2801844

[34] Chengxi Zang, Peng Cui, Chaoming Song, Christos Faloutsos, and Wenwu Zhu. 2017. Quantifying Structural Patterns of Information Cascades. In *Proceedings of the 26th International Conference on WWW Companion*. 867–868.

[35] 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 *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1355–1364.

[36] 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.