Come-and-Go Patterns of Group Evolution: A Dynamic Model

Tianyang Zhang[1], Peng Cui[1], Christos Faloutsos[2], Yunfei Lu[1], Hao Ye[3], Wenwu Zhu[1], Shiqiang Yang[1]
[1]Department of Computer Science and Technology, Tsinghua University
[2]Computer Science Department, Carnegie Mellon University
[3]Tencent Corporation, Shenzhen, China
zhangty09@foxmail.com, cuip@tsinghua.edu.cn, christos@cs.cmu.edu
luyf12@mails.tsinghua.edu.cn, dariaye@tencent.com, wwzhu@tsinghua.edu.cn, yangshq@tsinghua.edu.cn

ABSTRACT

How do social groups, such as Facebook groups and Wechat groups, dynamically evolve over time? How do people join the social groups, uniformly or with burst? What is the pattern of people quitting from groups? Is there a simple universal model to depict the come-and-go patterns of various groups?

In this paper, we examine temporal evolution patterns of more than 100 thousands social groups with more than 10 million users. We surprisingly find that the evolution patterns of real social groups goes far beyond the classic dynamic models like SI and SIR. For example, we observe both diffusion and non-diffusion mechanism in the group joining process, and power-law decay in group quitting process, rather than exponential decay as expected in SIR model. Therefore we propose a new model COMENGO, a concise yet flexible dynamic model for group evolution. Our model has the following advantages: (a) unification power: it generalizes earlier theoretical models and different joining and quitting mechanisms we find from observation. (b) succinctness and interpretability: it contains only six parameters with clear physical meanings. (c) accuracy: it can capture various kinds of group evolution patterns preciously and the goodness of fit increase by 58% over baseline. (d) usefulness: it can be used in multiple application scenarios such as forecasting and pattern discovery.

CCS Concepts

•**Information systems** → **Data mining;**

Keywords

Group Evolution; Dynamic Model; Temporal Patterns

1. INTRODUCTION

Forming social groups is an inherent human nature and the evolution of social groups is an essential mechanism in society. Understanding how these groups take shape and evolve over time is one of the central building blocks of modeling and understanding various social phenomena such as information spreading [13, 4], social-tie formation [10, 25], and social cooperations and competitions [9, 3]. Furthermore, studying the dynamics of groups also has rich practical use in various application scenarios. For example, the evolution of certain industry association can provide insight into the development trend of the business; the activities of online communities can become the vane of the popularity of new topics; and the detection of abnormal group growth can even help to handle the threat of online terrorist recruitment or find the trace of terrorist organization.

In literature, group evolution is a theme that runs through large parts of social science research [5]. Recently, some computational studies on social groups provide an profound foundation of the mechanism of group growth and also find several features that have significant influence on the group growth process [2, 11]. However, previous works still have some limitations and leave some fundamental problems unsolved.

The first fundamental problem is **how to model the temporal patterns of group evolution.** A comprehensive understanding of group evolution should contain both spatial dimension and temporal dimension. The spatial dimension studies how groups evolve over the underlying network while the temporal dimension studies how groups evolve over time. Several studies have been done in spatial dimension and they have found serval structure features influencing the evolution of group and the mechanism of how groups attract new member through the network [2, 22, 25, 11]. However, such feature based models can not capture the temporal patterns and reveal the inherent mechanism in temporal dimension.

The second fundamental problem is **how to model the quit process of group.** Group evolution process is the resultant of both come and go: a group attracts new members through diffusion or non-diffusion mechanisms [11] and meanwhile, members in the group may also leave the group. Most studies mainly focus on the growth mechanism but ignore the fact that, in most situations, groups also have quit mechanism, for both online and offline groups. Although some researches find the importance of member mobility in group evolution [7], there is no explicit model for the quit mechanism of groups. Compared to our understanding about how people join groups, how people quit groups remains quite unclear.

PLDI '13 June 16–19, 2013, Seattle, WA, USA

DOI: 10.475/123_4

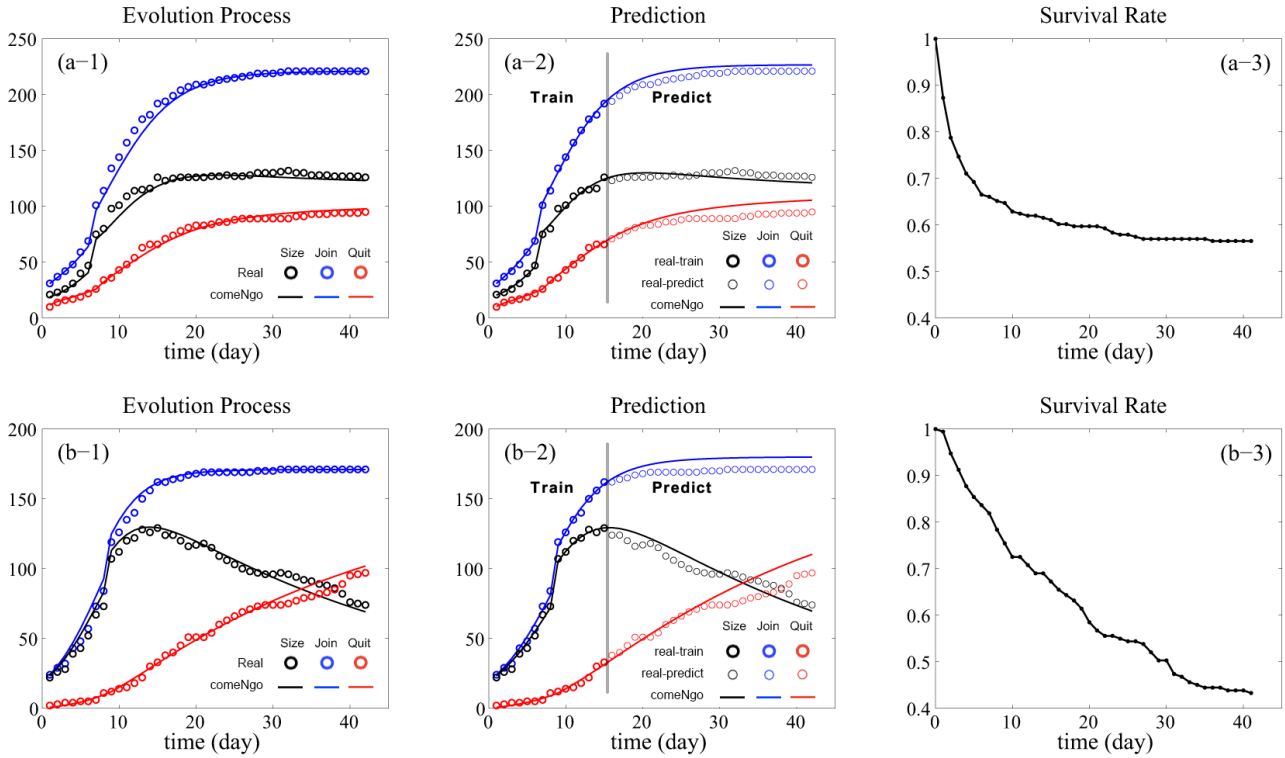

Figure 1: (a-1) and (b-1) The evolution process of two groups and in 42 days, the dots represent real data and the lines are the results of our model. (a-2) and (b-2) The prediction result of our model. We use the data of first 15 days to train the model. (a-3) and (b-3) Survival rate of holding time of these two groups.

In order to solve the two fundamental problems, we analyze a large real dataset with 103,548 groups from Wechat, the largest social communication network in China. Note that this is the first-ever large-scale dataset providing detailed logs of users joining and quiting social groups.. We observe rich complexities in the temporal patterns of group evolution. In fig 1 we showcase the group evolution process of two online groups over a time period of 42 days. From Figure 1.1, we can see different growth patterns with respect to group size of the two groups: one group gradually grows and then stays relatively stable while the other group shows continuous decline after a quick growth. We further look into the joining and quitting processes respectively. In quitting process, we observe different patterns of quit rates in different groups, as shown in Figure 1.3. The resulted holding time (the time interval between join group and quit group) distributions for different groups follows exponential decay (as expected in SIR, only account for 15% of all groups), power-law decay or inbetween. However, there is no published model to deal with the latter two cases, which account for the majority of groups. Moreover, we observe both diffusion and non-diffusion mechanism in joining process, while non-diffusion mechanism is beyond the capability of traditional dynamic models such as SI and SIR. Based on the observations, we propose a unified dynamic model for the group evolution process with temporal data. Our proposed model has the following appealing advantages:

- **Unification power:** It is a general model that includes previous models as special case and also embodies new mechanisms.
- **Succinctness and interpretability:** It uses only six parameters and every parameter has a clear physical meaning.
- **Accuracy:** It can capture the temporal patterns of different groups precisely as shown in fig 1.1. We achieve an improvement of **58%** in goodness of fit in a real dataset with more than 100 thousands groups.
- **Usefulness:** It is useful in multiple application scenarios such as forecasting and pattern discovery. Fig 1.2 is a showcase. With the early stage information of the group evolution process, our model can predict the remained evolution process accurately.

The rest of the paper goes as follows: Section 2 presents the fundamental concepts about dynamic models and Section 3 the proposed model. Section 4 shows our experimental results on a large-scale real dataset. We conclude in Section 5.

2. RELATED WORK

With the rapid growth of different kinds of online social network, we have more source and opportunity to study the structure and dynamics of social groups. In recent years, a lot of studies have been done on online group evolution patterns.

As one of the first empirical investigations on online groups, Backstrom, et al. study how structural features influence group formation and evolution in large social networks [2]. They reveal an important mechanism that group grows through the ties its members have to individuals outside the group. This mechanism is similar to the 'word-of-mouth' mechanism in diffusion process [21]. Several studies also focus on the influence of structural properties of groups in various underlying networks such as online games [8], mobile communication network [24, 25], co-authorship network [25, 14] and online social networks [22, 15]. Lin, et al. propose a probabilistic framework to analyze the evolution of community structures on a dynamic network [19]. Kairam, et al. conduct further research to reveal the more complex group evolve mechanism by making a conceptual difference between diffusion and non-diffusion growth in groups [11]. They find that non-diffusion mechanism also plays an important role in group evolution process.

As noted earlier, the previous works provide a profound explanation of how group grows through the network but do not answer the equally fundamental question that how group grows over time. Another limitation is that the do not reveal and model the mechanism of quitting the group. To both utilize the achievements and overcome the limitations of the previous works, we propose a dynamic model that captures the temporal patterns of both join and quit process.

3. PROPOSED METHOD

In this section we present our proposed method and analyze it.

First of all, our model should be a dynamic model that can present both temporal dimension and spatial dimension of group evolution. Besides, our model tries to capture the following behaviours, that we observed from our real data.

- P1: dependence mechanism: power-law holding time distribution in quit process.
- P2: non-diffusion mechanism: impulses in join process

To handle P1, we assume that the recoverability of a node may decay with the holding time exponentially, which is discussed in subsection 3.2. To handle P2, we import external shocks in our model, which is discussed in subsection 3.3.

We describe our model in steps, adding complexity, and we start with preliminaries..

3.1 Preliminaries

Dynamic models are widely used to describe the mechanism of temporal process such as epidemic spreading [1, 27], information diffusion [17, 20, 28], network evolving [16, 6]. In this subsection, we present the fundamental concepts about dynamic model. We start from one of the most basic dynamic models in spreading process: epidemiology models. We consider epidemic models as the baseline because they can capture the 'word-of-mouth' diffusion mechanism and can deal with the quit mechanism.

Susceptible-Infected model. The most basic epidemiology model is 'Susceptible-Infected'(SI) model. Each object/node is in one of two states - Susceptible(S) or Infected(I). Each infected node attempts to infect each of its neighbors independently at a constant rate β, which reflects the strength of the virus. Once infected, each node stays infected forever. The dynamic equation of the basic SI model is:

$$\frac{dI(t)}{dt} = \beta * (N - I(t))I(t) \tag{1}$$

where N is the total amount of obejcts/nodes and we assume that the underlying network is a clique of N, the time t is considered continuous, dI/dt is the derivative, and the initial condition I(0)>0 need to be given.

Susceptible-Infected-Recovered model. Another basic epidemiology model is an extension of SI model called 'Susceptible-Infected-Recovered'(SIR) model [12]. It allows Infected node to get recovered. Each object/node is in one of three states - Susceptible(S), Infected(I) or Recovered(R). The infect process is similar to SI model with a constant infect rate β while each infected node may get recovered at a constant rate γ. Once get recovered, the node will neither infect other nodes nor get infected by other nodes. The dynamic equation of the basic SIR model is:

$$\begin{aligned} \frac{dI(t)}{dt} &= \beta * (N - I(t) - R(t))I(t) - \gamma I(t) \\ \frac{dR(t)}{dt} &= \gamma I(t) \end{aligned} \tag{2}$$

where the initial condition R(0)=0.

In SIR model, the recover rate $\gamma(\tau)$ is a constant γ and we can derive the holding time follows an exponential distribution, that is,

THEOREM 1. *Let $f(\tau)$ be the probability density function of holding time τ in SIR model, then*

$$f(\tau) = \gamma e^{-\gamma\tau}$$

PROOF. By definition of $f(\tau)$,

$$f(\tau) = -\frac{dPr\{T > \tau\}}{d\tau} \tag{3}$$

By definition of $\gamma(\tau)$,

$$f(\tau) = Pr\{T > \tau\}\gamma(\tau) = \gamma Pr\{T > \tau\} \tag{4}$$

Take the derivative of Eq.5 with respect to τ and combine with Eq.4, we can derive that,

$$\frac{df(\tau)}{d\tau} = -\gamma f(\tau)$$

Solve this differential equation with initial condition $f(0) = \gamma$ and we can derive that,

$$f(\tau) = \gamma e^{-\gamma\tau}$$

□

This theorem demonstrate that the holding time distribution of SIR model follows an exponential distribution with an exponent of $-\gamma$. However, as is shown in Fig. 2, the holding time distribution of real data follows a power-law distribution, so the recover function $\gamma(\tau)$ should be a function decays verses holding time τ.

The above are the parameters of the base model and we will model the group evolve process as follows:

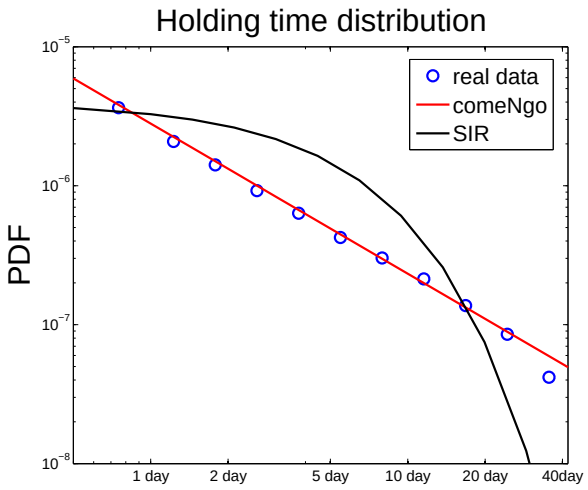

Figure 2: holding time distribution.

- n initial users found a group at birth time 0
- the group occasionally attracts other users in the population N
- users in the group may also quit the group

Note that the n initial users are also included in the population N and we assume that users who quit the group will not add it again, which is similar to SIR model.

Next we present our COMENGO model, which is a concise yet flexible dynamic model with several desirable properties.

3.2 Base model

3.2.1 *Dynamics equation*

A group evolve process can be seen as the interaction of two processes: join process and quit process. We also build our model on these two processes. Let $J(t)$ be the cumulative number of people who ever joined the group before time t, and $Q(t)$ be the cumulative number of people who quit the group before time t. Let $I(t)$ be the number of people in the group at time t, so the group evolve process $I(t)$ can be defined as follows,

$$I(t) = J(t) - Q(t) \tag{5}$$

So our task is to find the dynamics equation for $J(t)$ and $Q(t)$. For join process $J(t)$, we only consider the word-of-mouth mechanism in base model. We assume that each member in the group occasionally attracts other users to join the group and the possibility is determined by the attractiveness of the group β. So the dynamic equation of join process is as follows,

$$J'(t) = \frac{dJ}{dt} = \beta(N - J(t))I(t) \tag{6}$$

For quit process $Q(t)$, we need to consider the holding time distribution problem. As is proved in Theorem. 1, a constant recover function $\lambda(\tau)$ will lead to a exponential holding time distribution. In our model, we add an power-law decay with an exponent α to the recover function as follows,

$$\gamma(\tau) = \gamma_0 \tau^{-\alpha} \tag{7}$$

We will prove that the $\gamma(\tau)$ can generate both power-law and exponential holding time distribution depending on α. We first derive the probability density function $f(\tau)$ of holding time distribution as follows,

LEMMA 1. *Let $f(\tau)$ be the probability density function of holding time τ in* COMENGO *model, then*

$$f(\tau) = c\tau^{-\alpha} \exp(\frac{\gamma_0 \tau^{1-\alpha}}{\alpha - 1}) \tag{8}$$

PROOF. By definition of $f(\tau)$,

$$f(\tau) = -\frac{dPr\{T > \tau\}}{d\tau} \tag{9}$$

By definition of $\gamma(\tau)$

$$f(\tau) = Pr\{T > \tau\}\gamma(\tau) \tag{10}$$

this is,

$$Pr\{T > \tau\} = \frac{1}{\gamma_0}\tau^{\alpha} f(\tau) \tag{11}$$

Take the derivative of Eq.9 with respect to τ and combine with Eq.11, we can derive that,

$$-f(\tau) = \frac{\alpha}{\gamma_0}\tau^{\alpha-1} f(\tau) + \frac{1}{\gamma_0}\tau^{\alpha}\frac{df(\tau)}{d\tau}$$

Solve this differential equation, we can get that,

$$f(\tau) = c\tau^{-\alpha} \exp(\frac{\gamma_0 \tau^{1-\alpha}}{\alpha - 1})$$

where c is a constant decided by the initial condition. □

Now we prove that the holding time distribution can be both power-law tail and exponential tail depending on α.

THEOREM 2. *when $\alpha > 1$,*

$$\lim_{\tau\to\infty} f(\tau) \propto \tau^{-\alpha}$$

PROOF.

$$\lim_{\tau\to\infty} \frac{f(\tau)}{\tau^{-\alpha}} = \lim_{\tau\to\infty} c \cdot exp(\frac{\gamma_0 \tau^{1-\alpha}}{\alpha - 1})$$

$$\because \alpha > 1 \quad \therefore \lim_{\tau\to\infty} \tau^{1-\alpha} = 0, \quad \lim_{\tau\to\infty} \exp(\frac{\gamma_0 \tau^{1-\alpha}}{\alpha - 1}) = 1$$

$$\therefore \lim_{\tau\to\infty} \frac{f(\tau)}{\tau^{-\alpha}} = c, \quad \lim_{\tau\to\infty} f(\tau) \propto \tau^{-\alpha}$$

□

THEOREM 3. *when $\alpha < 1$,*

$$\lim_{\tau\to\infty} f(\tau) \propto e^{-\gamma\tau^{1-\alpha}}$$

PROOF.

$$\lim_{\tau\to\infty} \frac{\log f(\tau)}{\tau^{1-\alpha}} = \lim_{\tau\to\infty} \frac{\log c}{\tau^{1-\alpha}} - \alpha\frac{\log \tau}{\tau^{1-\alpha}} + \frac{\gamma_0}{\alpha - 1}$$

$$\because \alpha < 1 \quad \therefore \lim_{\tau\to\infty} \tau^{\alpha-1} = 0, \quad \lim_{\tau\to\infty} \tau^{\alpha-1}\log\tau = 0$$

$$\therefore \quad \lim_{\tau\to\infty} f(\tau) \propto e^{-\gamma\tau^{1-\alpha}}$$

□

As a result,the tail of the holding time distribution generated by our model mainly depends on the decay parameter α. When $\alpha > 1$, the holding time distribution will be power-law tailed while when $\alpha < 1$, it will have an exponential decay, and when $\alpha \to 0$, it will become an exact exponential distribution.

We also show the result of holding time distribution in Fig.2. The blue dot represents the holding time distribution of all groups and the red line is the result of our model. We can see from the figure that it can approximate the real data quit well. This evidence further demonstrates the correctness of our model.

With the appropriate recover function $\gamma(\tau) = \gamma_0 \tau^{-\alpha}$, we can derive the dynamics equation for the quit process as follows,

$$Q'(t) = \frac{dQ}{dt} = \int_0^t J'(x) f(t-x) dx \tag{12}$$

where $f(\tau)$ is defined in Eq.8.

Now we have the dynamics equation for join process (Eq.7) and quit process (Eq.15), in the next section, we will propose our dynamic model in discrete time based on the equation.

3.2.2 comeNgo-Base

Let $I(n)$ be the number of people in the group at time-tick n, let $\Delta J(n)$ be the number of people join the group at time-tick n, and let $\Delta Q(n)$ be the number of people quit the group at time-tick n. Our base model is governed by the equations

Model 1 (comeNgo-Base). *Our base model is generated by the equations*

$$\begin{aligned} I(n+1) &= I(n) + \Delta J(n) - \Delta Q(n) \\ \Delta J(n+1) &= \beta(N - J(n))I(n)/N \\ \Delta Q(n+1) &= \sum_{t=0}^{n} \Delta J(n) \cdot f(n+1-t) \end{aligned} \tag{13}$$

where,

$$f(\tau) = \gamma_0 \tau^{-\alpha} \exp(\frac{\gamma_0(\tau^{1-\alpha} - 1)}{\alpha - 1})$$

and initial conditions:

$$\Delta J(0) = J_0, \Delta Q(0) = 0$$

and also, by definition

$$J(n) = \sum_{t=0}^{n} \Delta J(t), \quad Q(n) = \sum_{t=0}^{n} \Delta Q(t)$$

Justification of the model. There are 4 parameters in our base model, $\{\beta, \gamma, \alpha, N\}$. An advantage of our dynamic model is that all the parameters have clear and explainable physical meaning, which means our model is more than fitting the data. Following is a brief explanation to each parameter:

- $\boldsymbol{\beta}$ indicates the attractiveness of the group. It determines how fast the group will attract new users and grow larger.
- $\boldsymbol{\gamma_0}$ indicates group members' short time satisfaction degree of the group. The higher the value is, the higher percentage of users will feel boring and quit the group soon.
- $\boldsymbol{\alpha}$ indicates group members' long time dependence on the group. If α is high, it means that the users who have been in the group for some time may become dependent on the group and are unlikely to quit anymore. In contrast, if α is low, group members are less likely to develop the long term interest in the group and the will gradually quit the group, sooner of later.
- $\boldsymbol{N}$ is the population of all potential members of the group. Although N may change over time, but it is relatively steady in most condition. So we set it to a constant in our model for brief.

We should also note that the initial condition in our model $\Delta J(0) = J_0$ needs to be given. It is the initial size of the group when found. Since group evolves mainly through diffusion mechanism, the initial size will greatly influence the growth speed at the beginning.

3.3 With Non-diffusion Growth

Although groups grow mainly through influence between nodes, non word-of-mouth mechanism may also plays an important part in group evolution process. Let's take the evolution of a new association in university for example. In the one hand, the founders and members may try to attract their friends into the association. This is 'word-of-mouth' mechanism and such mechanism will last all the time. In the other hand, they may also conduct membership recruitment meeting to attract new members. Such activity may attract a lot of new members in a short time, causing a burst like an external shock. Such non-diffusion mechanism may have great impact on the whole evolve process so we need to reflect it in our model.

We add the external shocks in the join process to model the non-diffusion growth. The dynamics equation of the join process is as follows,

$$J'(t) = \frac{dJ}{dt} = \beta(N - J(t))I(t) + \sum \lambda_i \delta(t - t_i) \tag{14}$$

where $\delta(t - t_i)$ is dirac delta function that indicates an impulse at time t_i, and λ_i is the strength of the shock.

So we propose our model as follows,

Model 2 (comeNgo).

$$\begin{aligned} I(n+1) &= I(n) + \Delta J(n) - \Delta Q(n) \\ \Delta J(n+1) &= \beta(N - J(n))I(n)/N + \sum_{i=1}^{k} \lambda_i \cdot \mathbf{1}_{\{t_i\}}(n) \\ \Delta Q(n+1) &= \sum_{t=0}^{n} \Delta J(n) \cdot f(n+1-t) \end{aligned} \tag{15}$$

where $\mathbf{1}_{\{t_i\}}(n)$ *is an indicator function, by definition,*

$$\mathbf{1}_A(x) := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

Justification of the model. The model is identical to COMENGO-BASE with the addition of the impulse term $\lambda_i \delta(t - t_i)$. This term captures the effect of non-diffusion mechanism in group evolution process. The impulse term indicates an external shock occurs at time t_i and attracts λ_i users to the group. Following is a brief explanation to the parameters of this term:

- $\boldsymbol{t_i}$ is the time when the external shock, such as a recruitment campaign, occurs. Such campaign can also last for some time but we regard it as an impulse because the duration is actually quite short compared with the duration of the word-of-mouth effect.
- $\boldsymbol{\lambda_i}$ is the effect strength of the impulse. λ_0 users will be recruited into the group and after that they can attract other users through word-of-mouth mechanism.

3.4 Parameter Learning

Our base model consists of four parameters: $\boldsymbol{\theta} = \{\beta, \gamma, \alpha, N\}$. Given two real time sequences X(n) and Y(n), X(n) is the cumulative number of people who ever joined the group before time-tick n and Y(n) is the cumulative number of people who ever quit the group before time-tick n, $(n = 1, ..., n_d)$. Note that the initial condition J_0 means the initial size of the group, so we can directly get from the data. To learn the parameters of the base model, we use Levenberg-Marquardt (LM) [18] to minimize the sum of errors:

$$\min_{\theta} D(X, Y, \theta) = \sum_{n=1}^{n_d} (X(n) - J(n))^2 + (Y(n) - Q(n))^2 \quad (16)$$

Our model with non-diffusion mechanism consists two more parameters in impulse term, location parameter t_i and strength parameter λ_i. We first determine the location parameter by detecting external shock larger than predefined threshold and then learn the strength parameter together with other parameters using Levenberg-Marquardt(LM) method [18].

4. EXPERIMENTS

To evaluate the effectiveness of COMENGO model, we conduct experiments on a real dataset. First we provide a brief data description in Section 4.1. In Section 4.2, we show how well we match the real data. In Section 4.3, we demonstrate the strong predicting power of our model. In section 4.4, we analyze the distribution of parameters and illustrate how to recognize different group evolving patterns.

4.1 Data Description

We conduct experiments on a real large-scale social network dataset - **Wechat**. Wechat is the most widely used social network and messaging service in China with more than 600 million monthly active users.

Group is one of the most important feature in Wechat. Every user can found groups to chat together. According to our statistics, more than one million groups are founded everyday. There are mainly two ways to add new members to the group. The usual way is by invitation. Every member in the group can invite their friends to the group, which is a typical diffusion (word-of-mouth) mechanism. The other way is by scanning the QR code. Users can generate a QR code for their group and promote the group by publishing the QR code in open places. For example, the group of an association can print the QR code on their posters and leaflets so that people who are interested in the association can scan it and join the group. This is a non-diffusion mechanism.

Name	Value
Group Number	103548
Time Duration	42 days
Total Join Records Number	10675984
Average Number of Join Records	103.10
Total Quit Records Number	5713719
Average Number of Quit Records	55.18

Table 1: Dataset Description

We use the logs of wechat group as our data. From the groups established in November 20th, 2015 which contain at least 40 join records before January 1st, 2016, we sampled 103,548 groups as our experiment data. For each group we have the following three kinds of records. Table 1 summarizes the statistics of the dataset used for this study.

- **Group Founding Records** $\mathcal{G}$: It consists of all the group foundation records (C, T) for each of the sampled groups. Each record means group C is founded at time T.
- **Group Joining Records** $\mathcal{J}$: It consists of all the group joining records (u, C, T) for each of the sampled groups. Each record means user u joined group C at time T.
- **Group Quitting Records** $\mathcal{Q}$: It consists of all the group quitting records (u, C, T) for each of the sampled groups. Each record means user u quitted group C at time T.

This is the first dataset for large scale social groups with detailed log of both joining and quitting behaviours. The dataset provides valuable opportunity to get insight about the temporal dynamics of group evolution mechanism. For privacy issues, the dataset is fully anonymized and all data are collected according to the terms and conditions of Wechat.

4.2 Discovering and Matching Group Evolution Patterns

In this section, we demonstrate how well our model matches the real data. First, we give the formulation of group evolving process for the real data. As is mentioned in section 3.1.1, a group evolving process can be regarded as the integral of join process and quit process. For each group C_i in $\mathcal{G}$ founded at timestamp T_i, we can generate the join process sequence $X_i(n)$ and quit process sequence $Y_i(n)$ with a daily time tick as follows,

$$\begin{aligned} X_i(n) &= \|\{(u, C, T) \in \mathcal{J} | C = C_i, T - T_i < n * \Delta T\}\| \\ Y_i(n) &= \|\{(u, C, T) \in \mathcal{Q} | C = C_i, T - T_i < n * \Delta T\}\| \end{aligned} \quad (17)$$

where ΔT is the length of each time tick, which is one day in our experiment. Then the group size sequence $G(n)$ can be derived by

$$G_i(n) = X_i(n) - Y_i(n) \quad (18)$$

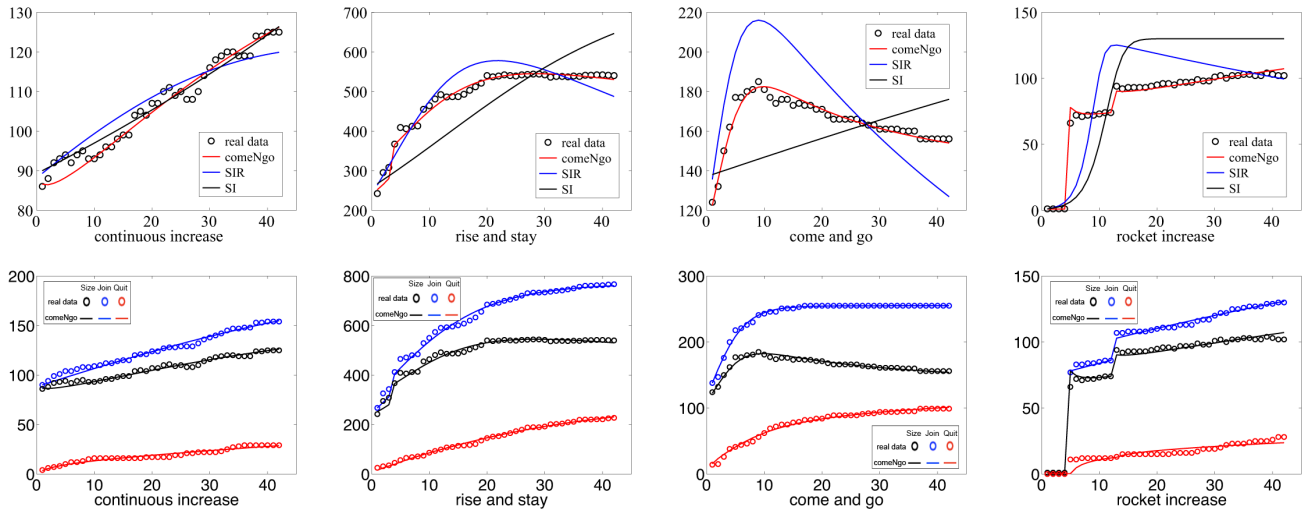

Figure 3: Fitting result of four kinds of patterns. Upper: Fitting result of group size. The dots present real data, the red line is the result of our model and blue for SIR model, black for SI model. Below: Fitting result of group size(black), join process(blue) and quit process(black).

Model	RMSE-I	RMSE-J	RMSE-Q
SI	18.17	-	-
SIR	11.07	5.36	9.12
comeNgo-Base	6.01	4.80	4.41
comeNgo	**4.64**	**2.63**	**3.81**

Table 2: Fitting accuracy of different models. Our comeNgo model outperforms all the baselines in fitting group size, join number and quit number.

Figure 3 shows the results of model fitting on Wechat dataset. We select 4 typical groups which represent different group evolving patterns. We show the original sequences $G(n)$(black dots) and fitting result of comeNgo $I(n)$(red line). We compare our result with SI model(blue line) and SIR model(black line). The second line of Figure 3 shows the fitting results of joining and quitting process. We can see that although the groups have quite different temporal patterns, our model can fit all the patterns very well, including the groups with external shocks (as shown in the fourth column of Figure 3). It demonstrates that our model is not only accurate but also has strong unification power.

We further evaluate the fitting accuracy using the root mean square error(RMSE) between estimated values and real values: $RMSE = \sqrt{\frac{1}{n_d}\sum^{n_d}(G(n) - I(n))^2}$. Besides group size sequences I(n), we also compare the fitting goodness of group joining sequences J(n) and group quitting sequences Q(n). We can see from the Table 2 that our model outperforms all the baselines. The fitting accuracy of our model increases by **58.08%** than baselines in fitting group size, **50.9%** in fitting joining process and **58.22%** in fitting quitting process.

4.3 Group Evolution Prediction

As a dynamic model, comeNgo model can be used to predict the whole process of group evolution. Here we consider a more practical problem: given the group evolving process in early stage, can we forecast the future evolution of the group? To answer the question, we designed two tasks, trend prediction and size prediction, to measure the predicting power of our model.

4.3.1 Trend Prediction

An important aspect of group evolving process is the evolving trend: will the group grow larger or begin to decline? For each group, we use the data of the first 10 days to train our model and use the learned parameters to predict the trend in the next 5, 15 and 30 days to demonstrate the predicting power in both short term and long term. Here we define the trend prediction problem as a classification problem: if the group size increases more than 20%, we regard it as a 'positive example' and if it decreases more than 20%, we regard it as a 'negative example'. Note that the external shocks are unpredictable to all models without external information, so we ignore the groups that have external shocks in prediction period. Besides, the number of positive and negative examples are balanced in this experiment.

We use the parameters of our model as features, and the distribution for each feature were standardized to have a mean of 0 and standard deviation of 1. Then we conduct Logistics Regression for the classification problem. We use 5-fold cross validation, that is, we randomly select 80% of the data as training data and the remaining 20% as testing data. The experiments are repeated for 5 times and the average performances are reported in Figure 4. We can see that our model outperforms all the baselines in three prediction tasks. In short term prediction, our accuracy is 81.13%, **14.3%** higher than baseline. In long term prediction, our accuracy is 80.10%, **14.9%** higher than baseline.

Table.3 shows the regression coefficients of the parameters of our model and we can find some meaningful information from them.

Parameters β, N and $\Sigma\lambda$ influence the dynamics of joining process. β and λ determine the growth speed through

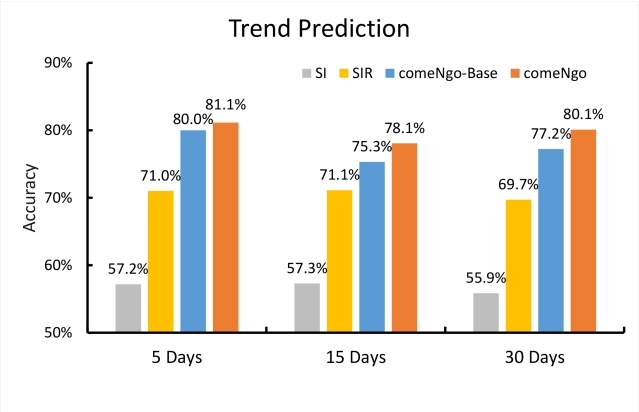

Figure 4: Prediction accuracy of the group growth trend in 5, 15 and 30 days. The accuracy of our model is 81.13%, 78.06% and 80.10% respectively.

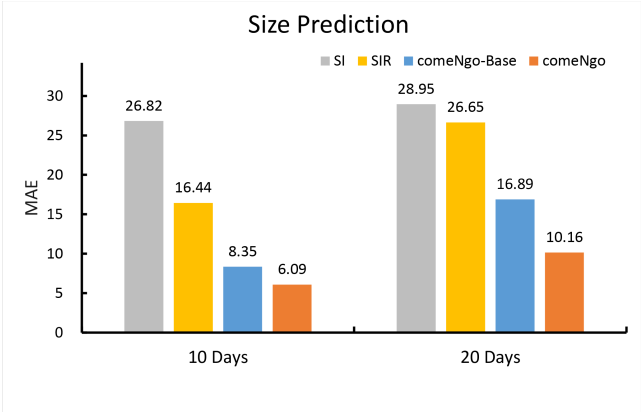

Figure 5: Prediction accuracy of the group size in 10 and 20 days. The average MAE of our model is 6.09 and 10.16 respectively.

Feature	5 Days	15 Days	30 Days
β	0.164**	—	-0.083*
γ	-1.52***	-1.13***	-1.05***
α	1.89***	1.95***	2.26***
N	0.318***	0.298***	0.295***
$\Sigma\lambda$	-0.598***	-0.652***	-0.731***
t_i	—	—	—

Table 3: Regression coefficients for comeNgo model predicting the trend in 5, 15 and 30 days after the 10th day(For this table, * p<0.01, p<0.005, *** p<0.001). Coefficients with p>0.05 are not reported.**

	10 Days		20 Days	
Model	MAE-J	MAE-Q	MAE-J	MAE-Q
SI	23.78	-	29.66	-
SIR	7.78	13.77	17.50	21.58
comeNgo-Base	8.36	6.39	20.73	12.02
comeNgo	**5.51**	**5.39**	**11.53**	**9.10**

Table 4: Results of predicting the join process and quit process. MAE-J and MAE-Q represent the average MAE of predicting cumulative join number and cumulative quit number in 10 and 20 days.

diffusion mechanism and non-diffusion mechanism respectively, while N determines the population of the users who can be reached by these mechanisms. A counterintuitive but interesting finding is that infection rate β and the external shock strength λ have negative coefficients for the long term growth trend. Meanwhile, the total population N of potential members shows significant positive coefficient for both short term and long term trend prediction. This finding indicates that growing breadth is more important than growing speed to the long term growth.

Parameters γ and α influence the dynamics of quit process. γ embodies more short term effect while α has more long term effect. We can find that the regression coefficient of α becomes larger when the prediction period is longer, in consist with its physical meaning. Also, γ plays an important role when predicting short term trend but decreases obviously when predicting long term trend. Considering that γ and α have the largest coefficients among all parameters, we can draw a conclusion that quit process has great impacts on group evolution trend.

4.3.2 *Size Prediction*

A more challenging and practical task to evaluate the prediction power of a model is the size prediction problem: if we only have the sizes of a group at different time ticks in early stage, can we predict the group size in future?

To demonstrate the predicting power of our model, we use the temporal data in the first 20 days to train our model and use it to predict the group size in the next 10 days and 20 days. We also ignore the groups with external shocks in prediction period, and there are 93660 groups left for the prediction problem.

We evaluate our prediction accuracy using the mean absolute error(MAE) between estimated values and real values: $MAE = |G(n) - I(n)|$. We also use SI and SIR as baseline models to predict the group size. The results are shown in Fig.5. We can see from the figure that our model achieves a significant improvement than baselines in both short term and long term. When predicting the group size in 10 days, the average MAE of our model is 6.09, **63.0%** better than SIR and in 20 days, the MAE is 10.16, also **61.9%** better than SIR.

Table.4 shows the predicting accuracy of group joining and quitting processes, measured by average MAE. SI model can only capture the joining process so it performs worst. SIR and comeNgo-Base model perform similar in predicting joining process, but our model performs much better in predicting quitting process. As is mentioned above, SIR model can only capture the non-dependence mechanism while our base model can capture both non-dependence and dependence mechanism. The improvement further demonstrates the importance of the dependence mechanism in quitting process.

4.4 Model Parameters Analysis

A specific advantage of our model is that all of the parameters have clear physical meanings. As a result, we can recognize different group evolving patterns through parameter analysis. In this subsection, we analyze the distribution

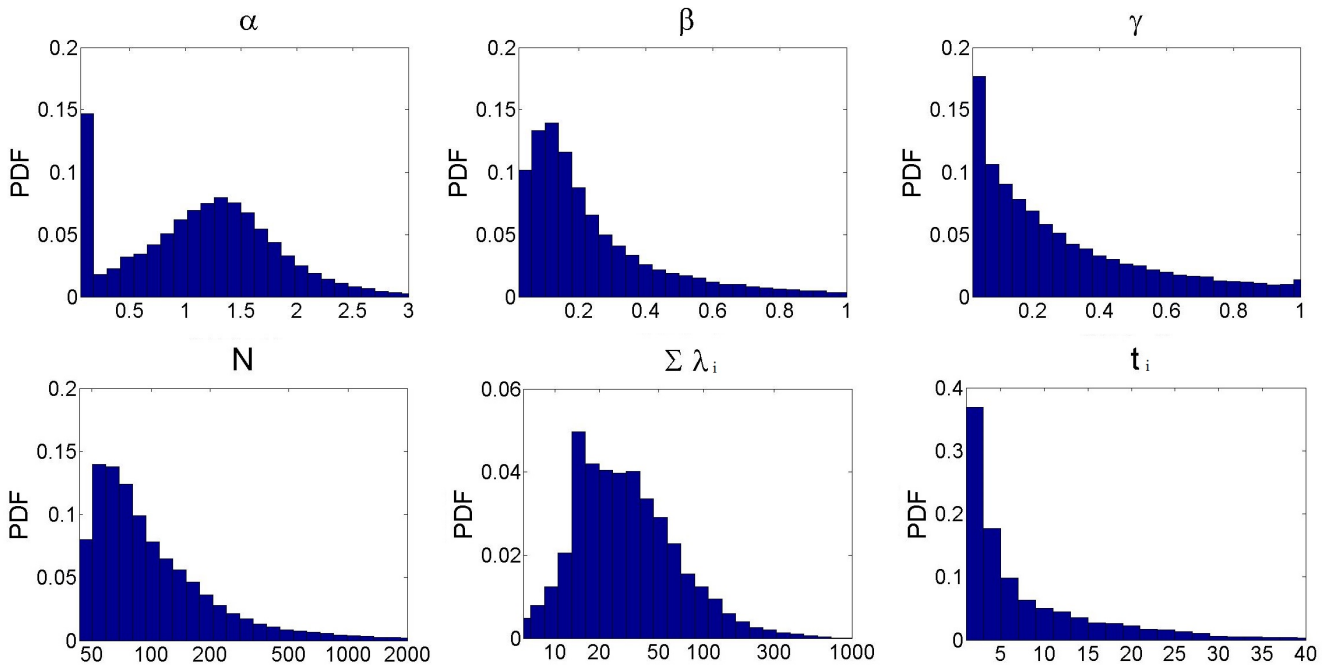

Figure 6: The distribution of six parameters in our model.

of the parameters and use them to recognize the temporal patterns in group evolution. Fig. 6 shows the distribution of six parameters in our model, and we get the following findings.

- $\boldsymbol{\alpha}$ **:** The distribution of α is a typical bimodal distribution with one peak near 0 and the other near 1.4. As we proved in Theorem 2 and 3, the clustering near the first peak represents the groups with exponential holding time distribution. In such groups, most people will quit by the non-dependence mechanism, and thus the group will gradually dies out. Such groups usually have weak ties between members. The other peak represents a typical power-law holding time distribution with an exponent near 1.5, which is consistent with many study in human dynamics [23, 26]. Groups with α much larger than 1.5 represents the groups with stable and strong relations, and thus members become dependent on the group shortly, such as phenomena are often observed in the the group of associations or organizations.

- $\boldsymbol{\beta}$ **:** It follows an natural exponential distribution and most groups have an infection rate less than 0.5.

- $\boldsymbol{\gamma}$ **:** It also follows an exponential distribution but it has an obvious peak near 0. This peak indicates that these groups have strong stickness and members barely quit the group. These groups are usually established by strong ties, such as classmates groups or working groups.

- $\boldsymbol{N}$ **:** The distribution of N follows a power-law (note that the x axis of N is presented in log scale) and the cutoff near 50 is resulted from our sampling strategy. The power-law decay indicates that the potential population N can be quite large and as we analyzed in Section 4.3.1, such groups are more likely to grow continuously.

- $\boldsymbol{\Sigma\lambda_i}$ **:** We find that 57.46% of groups do not have any external shocks in the group evolution, which are omitted in the figure. For the groups with external shocks, the non-diffusion mechanism are usually not strong,indicated by the peak near 20 and an power-law decay. This means that the diffusion mechanism plays more important roles for the evolution of most groups.

- $\boldsymbol{t_i}$ **:** Most external shocks occur at the early stage of the group, which is consistent with our intuition. For example, a new association usually holds many publicity campaigns to attract new members right after its establishment.

5. CONCLUSIONS

In this paper, we study the come-and-go patterns in group evolution process. We analyze the temporal patterns of group joining and quitting behaviours and revealed the dynamic mechanism behind these patterns. We propose COMEN-GO, a general, accurate and succinct dynamic model that is able to accurately model and comprehensively explain the temporal patterns. Our proposed model has the following appealing advantages:

- **Unification power:** It is a general model that includes previous models as special case and also embodies new mechanisms.

- **Succinctness and interpretability:** It uses only six parameters, what's more, every parameter has a clear physical meaning.

- **Accuracy:** It can capture the temporal patterns of different groups precisely, with an improvement of **58%** than baseline in goodness of fit.
- **Usefulness:** We showed how to use our model to do trend prediction, size prediction and recognize the temporal patterns by parameters analysis.

6. ACKNOWLEDGMENTS

This work was supported by National Program on Key Basic Research Project, No. 2015CB352300; National Natural Science Foundation of China, No. 61370022, No. 61531006, No. 61472444 and No. 61210008. Thanks for the research fund of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1314632 IIS-1408924 Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties.

7. REFERENCES

[1] R. M. Anderson, R. M. May, and B. Anderson. *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library, 1992.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM, 2006.

[3] R. S. Burt. The social structure of competition. *Explorations in economic sociology*, 65:103, 1993.

[4] D. Centola. The spread of behavior in an online social network experiment. *science*, 329(5996):1194–1197, 2010.

[5] J. S. Coleman and J. S. Coleman. *Foundations of social theory*. Harvard university press, 1994.

[6] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[7] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318. International World Wide Web Conferences Steering Committee, 2013.

[8] N. Ducheneaut, N. Yee, E. Nickell, and R. J. Moore. The life and death of online gaming communities: a look at guilds in world of warcraft. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 839–848. ACM, 2007.

[9] R. I. Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and brain sciences*, 16(04):681–694, 1993.

[10] M. S. Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

[11] S. R. Kairam, D. J. Wang, and J. Leskovec. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 673–682. ACM, 2012.

[12] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*, volume 115, pages 700–721. The Royal Society, 1927.

[13] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *World Wide Web*, 8(2):159–178, 2005.

[14] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila. Finding effectors in social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1059–1068. ACM, 2010.

[15] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470. ACM, 2008.

[16] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

[17] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. *SDM*, 7:551–556, 2007.

[18] K. Levenberg. A method for the solution of certain non–linear problems in least squares. 1944.

[19] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Analyzing communities and their evolutions in dynamic social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(2):8, 2009.

[20] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 6–14. ACM, 2012.

[21] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[22] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[23] J. G. Oliveira and A.-L. Barabási. Human dynamics: Darwin and einstein correspondence patterns. *Nature*, 437(7063):1251–1251, 2005.

[24] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[25] G. Palla, A.-L. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[26] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127, 2006.

[27] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *Reliable Distributed Systems, 2003. Proceedings. 22nd International Symposium on*, pages 25–34. IEEE, 2003.

[28] L. Yu, P. Cui, F. Wang, C. Song, and S. Yang. From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics. *arXiv preprint arXiv:1505.07193*, 2015.