Predicting Aggregate Social Activities using Continuous-Time Stochastic Process

Shu Huang[†], Min Chen[‡], Bo Luo[§], Dongwon Lee[†]

[†] College of IST, The Pennsylvania State University, University Park, PA, U.S.A.
[‡] Department of AMS, Johns Hopkins University, Baltimore, MD, U.S.A.
[§] Department of EECS, The University of Kansas, Lawrence, KS, U.S.A.

{shuang,dlee}@ist.psu.edu, minchen@jhu.edu, bluo@ku.edu

ABSTRACT

How to accurately model and predict the future status of social networks has become an important problem in recent years. Conventional solutions to such a problem often employ topological structure of the sociogram, i.e., friendship links. However, they often disregard different levels of *activeness* of social actors and become insufficient to deal with complex dynamics of user behaviors. In this paper, to address this issue, we first refine the notion of *social activity* to better describe dynamic user behaviors in social networks. We then propose a *Parameterized Social Activity Model (PSAM)* using continuous-time stochastic process for predicting aggregate social activities. With social activities evolving over time, PSAM itself also evolves and therefore dynamically captures the real-time characteristics of the current active population. Our experiments using two real social networks (Facebook and CiteSeer) reveal that the proposed PSAM model is effective in simulating social activity evolution and predicting aggregate social activities accurately at different time scales.

Categories and Subject Descriptors

H.1.0 [**Information Systems**]: Models and Principles; D.2.8 [**Models**]: Metrics—*data mining*

Keywords

Aggregate Social Activity, Continuous-time Stochastic Process

1. INTRODUCTION

The intensive and highly diversified communications among people lead to complex social network infrastructure. With the emergence of online social networking sites and online communities, a lot of research have been conducted to help understand the complex social phenomena[9, 29]. Among social networking research areas, an important topic is to understand and predict the dynamics of user behaviors and interactions [26, 10, 4], as well as the establishment and evolution of social relationship [17, 18].

In conventional social networking research, social networks are usually represented by graphs, where vertices denote *social actors* and edges denote the *social relationships*. Therefore, the notion of social network evolution is mostly defined as the expansion of the social network graph. Different approaches have been proposed to analyze and predict social network evolution and information diffusion. They can be roughly grouped into three categories: static network mining [30], microscopic evolution prediction [22, 24], and structural analysis [14, 32]. In static network mining, structural patterns, such as power-law distribution and small-world phenomenon, are discovered by mining snapshots of networks. In microscopic evolution prediction, various models are proposed to simulate the social network growth at the micro scope. In structural analysis, a variety of structural features are investigated to explain and predict the evolution of the social graph.

However, approaches that are purely based on graph structures could be biased when social actors demonstrate different levels of *activeness*. In particular, highly active social actors make more contribution towards the interaction, expansion and functionality of the online society, while inactive and less-active members only contribute as "silent" nodes in the graph. Therefore, we would argue that the notion of *social interactions* is an important indicator in social network analysis. Let us look at an example.

Example 1: Figure 1 shows two real-world social groups extracted from Facebook [28]. In terms of friendship links, their structural evolution processes are very similar. The final states of their friendship-network structures are shown in Figures 1 (a) and (b), which appear to be similar as well. Therefore, if we apply pure structure-based analysis on the two groups, they are expected to demonstrate similar behaviors (e.g., information diffusion). However, social interactions (e.g., wall postings) within the two groups turn out to be quite different, as exhibited in Figures 1 (c) and (d). That is, members in one group frequently communicate in Facebook, while the other one is very silent. □

As we can see, it is impractical to predict social network activities only using social connections (i.e., social relationships). As observed from Facebook data, a significant por-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10 ...$10.00.

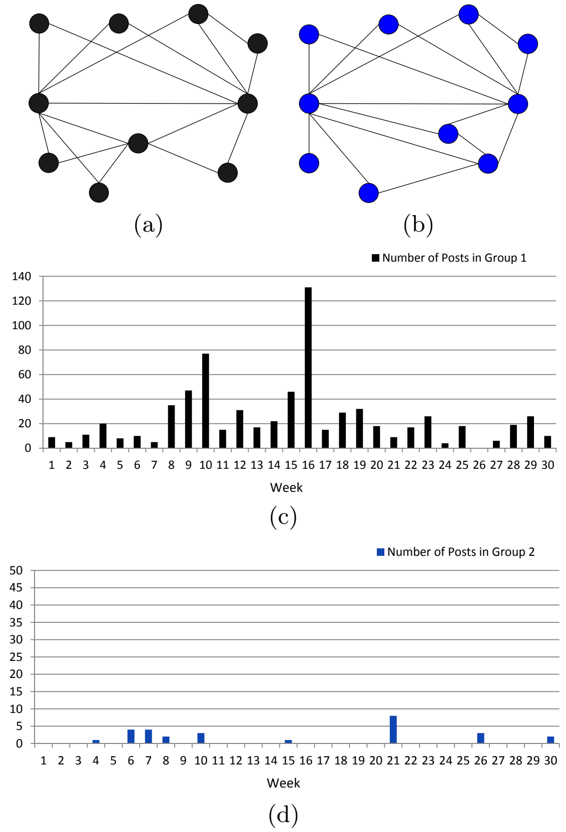


Figure 1: Example: two social groups with similar construction but different levels of interactions.

tion of the social network accounts are inactive, which causes purely structure-based evolution models to be less accurate. Moreover, compared with friendship links, *social activity* is a more meaningful indicator of *social dynamics* and *user behaviors* within social networks. In particular, social activities could be categorised into two types: *connective activities* and *interactive activities*. Connective activities (e.g., adding friends, subscripting, following) add new edges in the sociogram, and result in the expansion of social networks. Conventional research on social network evolution, which studies the development of the network structure, could be considered as the consequences of this type of activities. On the other hand, interactive activities, such as wall posting, poking and commenting, represent socialization interactions based on existing connections. This type of activities continuously contributes to the business value of social networking sites, and form the basis of online communities.

In this paper, we present a social activity predicting approach. By using a continuous-time stochastic process to simulate the social activity evolution, we address both the random impact from the environment and the tendency of evolution. At the same time, the model is parameterized with the current activity status, thus predictions are not purely based on historical information, but instead, based on the inherent characteristics and real-time observation of the social community. In addition, predictions made by our continuous-time stochastic model have an infinite future state space, i.e. numeric values instead of binary notions; therefore our model is advantageous in quantifying the activity evolution. Furthermore, the proposed approach could be applied on any arbitrary subset of users in a social network (to be elaborated), making the proposal very flexible. Finally, with comprehensive experiments on real-world social network data, the proposed method is shown to be effective and superior to baselines in predicting social activities.

Our major contributions in this paper are threefold:

(1) We introduce a social network evolution model based on social activity features, including both connective and interactive activities. This model better simulates and predicts the dynamics of social behaviors compared with purely structure-based models. The improved performance of the new model is demonstrated in experiments on real social networks.
(2) The parameters of the model also evolve with the day-to-day user activities in social communities. Therefore, the parameterized model is dynamic and accurately captures the active population in the network.
(3) The proposed model exploits aggregate interactions among a group of social network users. It is applicable to study an arbitrary subset of users, who are selected with diverse criterions. Such flexibility allows the model to be useful in a wide range of applications.

The rest of this paper is organized as follows: we introduce the problem and the background of continuous-time stochastic process in Sections 2 and 3. In section 4, we present the details of the parameterized social activity model. We show the experimental results in Section 5, discuss related works in Section 6, and finally conclude the paper in Section 7.

2. PROBLEM STATEMENT

2.1 Background

In this paper, we analyze and model a wide range of social behaviors in social networks, e.g., becoming a member, posting a message, adding a comment, inviting a friend, etc. We refer to such as *activity* in this paper. Further, an instance of such an activity is referred to as an *interaction*. For example, suppose we model an activity A_1 as "a member posting comments to others' walls in Facebook." If a member m_1 posts three comments on another member m_2's wall, it is said that m_1 and m_2 have three interactions.

User activities are very important in social network research, as they are basic indicators of user behaviors and social dynamics. Interactions among members also reveal the business value of social communities in multiple aspects. Meanwhile, the *realtime activeness* of social actors is also an important factor in the dynamic analysis and applications of social networks. For instance, in advertising, *ex-ante* knowledge to the status of the targeted social network will assist decision making and improve the profit; e.g. it is only worthwhile to advertise in active social groups that are potentially interested in the product/event. As another example, we can estimate the business potential of fans of college basketball teams by analyzing activities among fan groups, and inject advertisements when activities reach their peaks.

As we have introduced, social activities could be categorized into *connective activities* (i.e. establishing social connections) and *interactive activities* (i.e. information sharing and socialization). Despite much success in structural-based social network analysis, little work has been done in analyzing and predicting interactive activities such as wall postings. Meanwhile, in topology-based social network analysis, nodes and edges accumulate over time. However, this model does not explain the possible decrease of the active member population. Taking Facebook as an example, the active member population is not continuously increasing; instead, decrease and increase take place alternatively over time. Moreover, it is also challenging to simulate the influence of the environment on the social activity. The impact of the environment is usually random, which is hard to quantify and depict. Without quantifying the environment variables, it is difficult to predict the social activity accurately.

Intuitively, it is important to model both *individual activities* and *aggregate activities*. However, analysis of individual activities for all social network users may be impractical in many applications: (1) modeling and prediction at individual level could be biased due to the relatively large uncertainty and randomness. For instance, when we predict whether a user will reply to a message [13] or join a group [4], the likelihood is always low, i.e. users are always very unlikely to conduct any activity. Therefore, aggregate activity makes more sense as it represents collective user behaviors at a macro level, and the randomness on individuals offset each other. (2) It could be computationally too expensive to analyze every individual in the network. In particular, when we only expect aggregate results (e.g. the group of x users may post y messages in the next day), it is a waste to predict at individual level and then aggregate (e.g. predict the likelihood of posting a message for each user, and then compute the group-level aggregation).

In social network evolution, we observe that aggregate activities provide a good indication of social network status. Considering social activities from a macro scope, some activity features, such as the population of active members and the number of interactions, can serve to depict member activities. Meanwhile, activity features also imply future network evolution. Frequent member interactions usually indicate a growth of the social network, while low interactions may suggest a shrinking network. Therefore, social networks can be measured by incorporating a variety of activity features, and this measure can help predict future activities.

2.2 Problem Definition

In this paper, we aim at the problem of activity-based social behavior prediction. We emphasize the contribution of the active member interactions and exploit the activity features to predict social activities. Instead of the conventional sociogram that models users as nodes and friendship relationships as edges, we use the *social activity graph* that is formed by user interactions within a certain *time interval*. As we have introduced, types of interaction in a social network are highly diversified, e.g., adding a friend or posting on the wall in a friendship network, or collaborating on a publication in a co-authorship network. We formalize the concept of *social activity* (SA) as follows:

Definition 1 (Social Activity) *Given a subset M of the social network users (nodes), the social activities of M at time interval $[t_i, t_{i+1})$ is represented by a labeled graph $G(t_i) = (V_{t_i}, E_{t_i})$, where V_{t_i} is the vertex set and E_{t_i} is the edge set. Every vertex $v \in V_{t_i}$ corresponds to an individual who is involved in at least one activity on $[t_i, t_{i+1})$. An edge $e = \langle u, v \rangle, e \in E_{t_i}$ exists between a pair of vertex u and v if and only if u and v have at least one interaction of a preselected type of actions during the interval $[t_i, t_{i+1})$. The edge is labeled as the number of interactions between u and v during the interval.*

In the rest of the paper, we use *ISA* (interval-wise social activities) to denote social activity features within a time interval. With the definition, the membership is no longer permanent or cumulative. Only those who have activities at $[t_i, t_{i+1})$ are included in $G(t_i)$. Therefore, the membership and status of $G(t)$ is evolving over time.

The Problem. With the notions of social activity and social activity graph, we can ask many interesting questions. In particular, in this paper, we aim to answer the question: *how can we predict aggregate social activity at time t_{i+1} using social activities on the time up to t_i?*

Let N_{t_i} denote the social activity in a graph $G(t_i)$. Then, we can represent N_{t_i} in various ways: e.g. number of vertices $|V(t)|$ (i.e. number of active members), number of edges $|E(t)|$ (i.e. number of active pairs of users), or the number of interactions in the graph ($|I(t)|$), depending on the application. In the experiment, we examine all these three different measures in constructing and validating our parameterized social activity model.

Node selection. The proposed approach does not depend on the friendship network among users, hence, it could be applied on any arbitrary subset of users in a social network. In practice, nodes are selected based on the objective of the analysis/prediction, and the selection criteria could be structure-based (e.g. select the neighborhood of a seed user), attribute-based (e.g. select all the 2012 graduates of a school), content-based (e.g. select all users who have tweeted about New York Giants), or even activity-based (e.g. select the most active users in a community and analyze interactions between them). Theoretically, it is valid to select a subset of completely unrelated users. However, such selection makes no practical sense, since the interactions among selected users will be very sparse.

Solution overview. To quantify the influence of the environment and address the randomness, we make use of a continuous-time stochastic process, named *Wiener Process* (WP). With this tool, we derive a parameterized social activity evolution model, which takes into account the temporal change of the active population. By utilizing activity features, the proposed model dynamically fits the current network characteristics and accurately simulates the evolution of social activities. At last, the model generates a change rate distribution of the social activity status, based on which a prediction can be made.

3. PRELIMINARY: CONTINUOUS-TIME STOCHASTIC PROCESS

Due to the randomness in social dynamics, we seek to investigate the probability distribution of the ISA status N_t. As N_t is a continuous-time random variable, we simulate its evolution with a continuous-time stochastic process. A stochastic process is a process or system that evolves randomly but can be described by probability distributions. It is frequently used at macro scope to simulate the evolution of a random process. A continuous-time stochastic process is nowhere differentiable. To handle this property, stochastic calculus is developed with theoretical foundations and methods that are different from regular calculus.

In the evolution of a social network, N_t evolves with infinite future status, i.e. may take any numeric value. To address this property, we adopt the Wiener Process [16] as a component in our model. The Wiener Process (WP), also called Brownian Motion, is a fundamental and basic stochastic process. It is a key process to build other more complicated processes. WP is widely used in both pure and applied mathematics, in particular the Nobel-Prize-Winning result "Black-Scholes option pricing model" [16].

The definition of a Wiener Process $W(t)$ is as follows:

1. $W(0) = 0$.
2. $W(t)$ is almost surely continuous.
3. $W(t_2) - W(t_1)$ is a random variable for any $0 = t_0 < t_1 < t_2 < t_n$.
4. $W(t_2) - W(t_1)$ and $W(s_2) - W(s_1)$ are independent normal distributions that follows $N(0, \sqrt{t_2 - t_1}^2)$ and $N(0, \sqrt{s_2 - s_1}^2)$ respectively, when $[t_1, t_2]$ and $[s_1, s_2]$ do not overlap.

As a stochastic process, $W(t)$ is continuous but nowhere differentiable. Solving a continuous-time stochastic process involves some special properties, and the **quadratic variation** is the most important one. Denote $\prod = \{t_0, t_1, ...t_n\}$ as a partition of time interval $[0, T]$, and

$$||\prod|| = max_{j=0,\dots,n-1}(t_{j+1} - t_j),$$

The quadratic variation of $W(t)$ up to time T is defined as:

$$[W, W](T) = \lim_{||\prod||\to 0} \sum_{j=0}^{n-1} [W(t_{j+1}) - W(t)]^2 \tag{1}$$

where $0 = t_0 < t_1 < ... < t_n = T$. For all $T \geq 0$, $[W, W](T) = T$ almost surely [16]; in other words, *WP accumulates quadratic variation at rate one per unit time.* It can also be written as:

$$dW(t)dW(t) = dt \tag{2}$$

This property of quadratic variation distinguishes stochastic calculus from regular calculus. Consider a function $f(W(t))$, where $W(t)$ is a WP, according to equation (2), the differential of $f(W(t))$ is:

$$\begin{aligned} df(W(t)) &= f'(W)dW + \frac{1}{2}f''(W)dWdW \\ &= f'(W)dW + \frac{1}{2}f''(W)dt \end{aligned} \tag{3}$$

The equation above is different from what is obtained by regular calculus on regular functions. Any stochastic process that contains a WP inevitably inherits stochastic properties from the WP. As WP is a component in our parameterized evolution model, we will make use of stochastic calculus to generate the solution to our model.

By definition of WP, at time t_1, the expectation of future value of $W(t)$ equals the value of $W(t)$ at time t_1, i.e. $\mathrm{E}[W(t_2)|F_{t_1}] = W(t_1)$ for any $0 \leq t_1 < t_2$, F_{t_1} represents the information at time t_1. This can be interpreted to mean that WP has no tendency to increase or decrease in the future. Although a WP is effective to simulate the influence of the environment, it does not indicate the long-term evolution tendency. Taking Facebook social network as an example, as Facebook becomes more popular, it has attracted more members and they are likely to have more intensive social activity. However, WP can not reflect this long-term evolution tendency, because the expectation of $W(t)$ equals its value at the evaluating time.

At the same time, a WP does not address the magnitude of randomness in the process. By definition, all WPs are considered equally volatile, because the standard deviation of the random increment $W(t_2) - W(t_1)$ is always $\sqrt{t_2 - t_1}$. If two stochastic processes with different volatility are both modeled by a WP, then on the same time interval $[t_1, t_2]$, their models will have the same standard deviation, which is not desirable.

For these two reasons, a WP by itself is insufficient to describe the evolution of N_t. Therefore, based on a WP, we propose a parameterized stochastic process model to simulate the evolution of ISA.

4. THE PARAMETERIZED SOCIAL ACTIVITY MODEL

4.1 Overview

By incorporating the Wiener Process, we propose the **Parameterized Social Activity Model (PSAM)** to simulate the social activity evolution over continuous-time. In this model, N_t is the random variable with the following observations and assumptions: (1) it evolves over continuous time; (2) the process is shaped by both the tendency of evolution factors and random shock from the environment, e.g. users' behavior patterns and environment shock; (3) the evolution tendency of N_t on any time interval $[t_i, t_{i+1}]$ can be considered determined by the information available at time t_i, and is independent from previous history. According to these observations, a continuous-time stochastic process is a proper simulation for N_t.

In evolution of the social activity, influential factors include both random impact from environment and characteristics of the active population. All this information is reflected within activity features of the social community. Incorporating WP as a component, the randomness of the process is addressed and quantified. At the same time, a variety of activity features are extracted as predicting variables(will be elaborated later). As a summary of influential factors, they enable parameterization of the model and resolve the insufficiency of WP. By incorporating them as predicting vector, N_t is dynamically simulated and predicted.

In PSAM, no distribution of social activity is assumed in advance. Meanwhile, the selection of social community is not subject to any restriction. As a component of the model, the WP makes sure the simulation is continuous-time and predictions can be made on any time point. The parameterization guarantees dynamics of the model, therefore the latest influential factors are addressed and included when making predictions.

4.2 Activity Feature Extraction

To measure member activities, we construct ISA over successive time intervals and extract a list of activity features out of it. They include not only the features indicating activity level of members, but also those related to the topology

Table 1: Activity features extracted from ISA

Features	Description
$\lvert V_t \rvert$	Number of active members in $G(t)$
$\lvert CV_t \rvert$	Cumulative number of active members by time t
ΔV_t	Difference between $\lvert V_t \rvert$ and $\lvert V_{t-1} \rvert$
$\lvert I_t \rvert$	Number of interactions in $G(t)$
$\lvert CI_t \rvert$	Cumulative number of activities by time t
ΔI_t	Difference between I_t and I_{t-1}
$\lvert E_t \rvert$	Number of edges in $G(t)$
$\lvert CE_t \rvert$	Cumulative number of edges by time t
ΔE_t	Difference between E_t and E_{t-1}
AI_t	Average number of interactions per each person in $G(t)$
AR_t	Average number of friends per each person in $G(t)$
CC_t	Average clustering coefficient in $G(t)$
AL_t	Average length of the shortest pathes in $G(t)$
D_t	Diameter across all vertices of $G(t)$

change[4]. Table 1 shows all the 14 activity features generated from the Facebook and Citeseer social networks.

In our datasets, the first 9 features evolves significantly over time, therefore they are rescaled by logarithm to handle the very large values. The rescale of features does not change the predicting accuracy of our model.

As a lot of activity features are used, they may introduce high dimensional parameters in the model construction later on. To keep the model simple, proper techniques (e.g. principle component analysis) could be employed to reduce the data dimension. Since this is not the main focus of our study, we do not present more details here.

With all activity features obtained, the results are plugged into the evolution model as predictors. Let s_j denote the j-th dimension extracted from activity features, m denote the total number of dimensions used, then the predictor vector at time t_i is $\vec{\lambda}_{t_i} = [s_{1t_i}, s_{2t_i}, ...s_{mt_i}]^T$.

4.3 The Parameterized Evolution Model

Let $\vec{\lambda}_t$ represent the predictor vector, N_t denotes the status of ISA. To describe the evolution of N_t, we derive a parameterized stochastic process with a drift and diffusion as follows:

$$dN_t = \gamma(\vec{\lambda}_t)N_t dt + \sigma(\vec{\lambda}_t)N_t dW(t) \tag{4}$$

where t denotes the continuous time and $\vec{\lambda}_t$ is time-evolving. Both $\gamma(\vec{\lambda}_t)$ and $\sigma(\vec{\lambda}_t)$ are time-evolving parameters depending on $\vec{\lambda}_t$. $\gamma(\vec{\lambda}_t)N_t dt$ is the drift term, indicating the growth or shrinkage of active population. $\sigma(\vec{\lambda}_t)N_t dW(t)$ is the diffusion term, describing the uncertainty, e.g. the impact from the environment. $W(t)$ is a WP. Due to the existence of $W(t)$, N_t inherits the property of quadratic variation, and it is also continuous but nowhere differentiable.

Equation (4) not only describe the evolution tendency, but also address the magnitude of randomness in the process. As the coefficients σ and γ are parameterized with $\vec{\lambda}_t$, they allow the evolution model itself to evolve over time, and therefore reflect the latest characteristics of N_t.

The time-evolving coefficients make this model hard to work with. To solve this model, we make use of an approximated form. Denote $\prod = \{t_0, t_1, ...t_n\}$ as a partition of $[0, T]$ into intervals $[t_i, t_{i+1}]$, $0 \le i \le n-1$. Because the intervals are small, $\gamma(\vec{\lambda}_t)$ and $\sigma(\vec{\lambda}_t)$ can be treated as constant on each interval $[t_i, t_{i+1})$, as determined at time point t_i.

Let $0 \le \Delta t < t_{i+1} - t_i$ be the time increment, then on each interval $[t_i, t_{i+1})$, the equation (4) is approximated with the following format:

$$dN_{t_i+\Delta t} = \gamma(\vec{\lambda}_{t_i})N_{t_i+\Delta t}d\Delta t + \sigma(\vec{\lambda}_{t_i})N_{t_i+\Delta t}dW(\Delta t) \tag{5}$$

where $\gamma(\vec{\lambda}_{t_i})$ and $\sigma(\vec{\lambda}_{t_i})$ are kept constant for any Δt. For simplicity, let $\aleph$ denote $N_{t_i+\Delta t}$, then equation (5) equals:

$$d\aleph = \gamma(\vec{\lambda}_{t_i})\aleph d\Delta t + \sigma(\vec{\lambda}_{t_i})\aleph dW(\Delta t) \tag{6}$$

This yields:

$$\frac{d\aleph}{\aleph} = \gamma(\vec{\lambda}_{t_i})d\Delta t + \sigma(\vec{\lambda}_{t_i})dW(\Delta t) \tag{7}$$

Because $\aleph$ has the property of quadratic variation, from equation (3), we obtain:

$$d\ln\aleph = \frac{d\aleph}{\aleph} + \frac{1}{2}(-\frac{1}{\aleph^2})d\aleph d\aleph \tag{8}$$

It is then:

$$\frac{d\aleph}{\aleph} = d\ln\aleph + \frac{d\aleph d\aleph}{2\aleph^2} \tag{9}$$

Plugging equation (9) into equation (7) yields:

$$d\ln\aleph + \frac{d\aleph d\aleph}{2\aleph^2} = \gamma(\vec{\lambda}_{t_i})d\Delta t + \sigma(\vec{\lambda}_{t_i})dW(\Delta t) \tag{10}$$

Multiply equation (6) by itself, we obtain:

$$\begin{aligned} d\aleph d\aleph =& [\gamma(\vec{\lambda}_{t_i})\aleph d\Delta t + \sigma(\vec{\lambda}_{t_i})\aleph dW(\Delta t)]^2 \\ =& \gamma^2(\vec{\lambda}_{t_i})\aleph^2 d\Delta t d\Delta t + \sigma^2(\vec{\lambda}_{t_i})\aleph^2 dW(\Delta t)dW(\Delta t) \\ &+ 2\gamma(\vec{\lambda}_{t_i})\sigma(\vec{\lambda}_{t_i})\aleph^2 d\Delta t dW(\Delta t) \end{aligned} \tag{11}$$

Note that, by calculus, $d\Delta t d\Delta t = d\Delta t dW(\Delta t) = 0$. Therefore, the above equation yields:

$$d\aleph d\aleph = \sigma^2(\vec{\lambda}_{t_i})\aleph^2 dW(\Delta t)dW(\Delta t) \tag{12}$$

Furthermore, by *stochastic calculus* (equation (2)), equation (12) is then:

$$d\aleph d\aleph = \sigma^2(\vec{\lambda}_{t_i})\aleph^2 d\Delta t \tag{13}$$

Then plugging equation (13) into equation (10), we obtain:

$$d\ln\aleph = [\gamma(\vec{\lambda}_{t_i}) - \frac{1}{2}\sigma^2(\vec{\lambda}_{t_i})]d\Delta t + \sigma(\vec{\lambda}_{t_i})dW(\Delta t) \tag{14}$$

Integrating equation (14) on interval $[t_i, t_i + \Delta t]$, we have:

$$\ln\aleph = [\gamma(\vec{\lambda}_{t_i}) - \frac{1}{2}\sigma^2(\vec{\lambda}_{t_i})]\Delta t + \sigma(\vec{\lambda}_{t_i})W(\Delta t) + \ln N_{t_i} \tag{15}$$

With $\aleph = N_{t_i+\Delta t}$ recovered, equation (15) becomes:

$$N_{t_i+\Delta t} = N_{t_i}exp\{[\gamma(\vec{\lambda}_{t_i}) - \frac{1}{2}\sigma^2(\vec{\lambda}_{t_i})]\Delta t + \sigma(\vec{\lambda}_{t_i})W(\Delta t)\}$$

Now denote $\mu(\vec{\lambda}_{t_i}) = \gamma(\vec{\lambda}_{t_i}) - \frac{1}{2}\sigma^2(\vec{\lambda}_{t_i})$, then:

$$\frac{N_{t_i+\Delta t}}{N_{t_i}} = exp\{\mu(\vec{\lambda}_{t_i})\Delta t + \sigma(\vec{\lambda}_{t_i})W(\Delta t)\} \tag{16}$$

By definition of $W(t)$, its increment follows a normal distribution, i.e., $W_{t+\Delta t} - W_t \sim \text{Normal}(0, \sqrt{\Delta t}^2)$. Therefore, $\mu(\vec{\lambda}_{t_i})\Delta t + \sigma(\vec{\lambda}_{t_i})W(\Delta t) \sim \text{Normal}(\mu(\vec{\lambda}_{t_i})\Delta t, (\sigma(\vec{\lambda}_{t_i})\sqrt{\Delta t})^2)$. Then by definition of log-normal distribution, equation (16) implies:

$$\frac{N_{t_i+\Delta t}}{N_{t_i}} \sim \text{LogNormal}(\mu(\vec{\lambda}_{t_i})\Delta t, (\sigma(\vec{\lambda}_{t_i})\sqrt{\Delta t})^2) \tag{17}$$

To focus only on the ending points of time intervals, make $\Delta t \to t_{i+1} - t_i$. Since N_t is continuous, $N^-_{t_{i+1}} = N_{t_{i+1}}$, therefore $N_{t_i+\Delta t} = N_{t_{i+1}}$. Note that the time interval to sample N_t is constant (e.g. daily). Without loss of generality, we take $t_{i+1} - t_i$ as the base unit to measure time, i.e. $\forall i, t_{i+1} - t_i \equiv 1$. Thus we obtain:

$$\frac{N_{t_i+1}}{N_{t_i}} \sim \text{LogNormal}(\mu(\vec{\lambda}_{t_i}), (\sigma(\vec{\lambda}_{t_i}))^2) \tag{18}$$

Define $X_{t_i} = N_{t_i+1}/N_{t_i}$, the change rate of N_t, the probability density function of X_{t_i} is then:

$$\begin{aligned} & f_{t_i}(X_{t_i}; \mu(\vec{\lambda}_{t_i}), \sigma(\vec{\lambda}_{t_i})) \\ =& \frac{1}{x\sigma(\vec{\lambda}_{t_i})\sqrt{2\pi}} exp(-\frac{(lnx - \mu(\vec{\lambda}_{t_i}))^2}{2\sigma(\vec{\lambda}_{t_i})^2}) \end{aligned} \tag{19}$$

The values of parameters $\mu(\vec{\lambda}_{t_i})$ and $\sigma(\vec{\lambda}_{t_i})$ both evolve with $\vec{\lambda}_{t_i}$. To estimate the parameters, we adopt a linear relation between μ and $\vec{\lambda}_{t_i}$, and an exponentially linear relation on σ and $\vec{\lambda}_{t_i}$. That is:

$$\mu(\vec{\lambda}_{t_i}) = \vec{\alpha}^T \begin{bmatrix} 1 \\ \vec{\lambda}_{t_i} \end{bmatrix} = \alpha_0 + \sum_{j=1}^{m} \alpha_j s_{jt_i} \tag{20}$$

$$\sigma(\vec{\lambda}_{t_i}) = exp(\vec{\beta}^T \begin{bmatrix} 1 \\ \vec{\lambda}_{t_i} \end{bmatrix}) = exp(\beta_0 + \sum_{j=1}^{m} \beta_j s_{jt_i}) \tag{21}$$

where $\vec{\alpha} = [\alpha_0, \alpha_1, ..., \alpha_m]^T$ and $\vec{\beta} = [\beta_0, \beta_1, ..., \beta_m]^T$ are the coefficient vectors of μ and σ respectively.

$\vec{\alpha}$ and $\vec{\beta}$ can be estimated with historical records, thus the parameterized model is partially determined. With a new ISA and activity features, the predictor vector $\vec{\lambda}_{t_i}$ takes new values, then a new distribution of X is generated by equation (19) to describe the evolution of N_t.

4.4 Prediction

Given a new ISA at time t_i, to predict its evolution, we start with the activity feature extraction. By applying the activity features to the parameterized evolution model, the current distribution of X can be generated. Then, according to the distribution, we calculate the upper and lower bound of the $1 - \alpha$ confidence interval. Based on the evolution model, we propose to predict the change rate at t_i as follows:

$$x_{t_i} = \frac{exp(\mu_{t_i} - \sigma_{t_i} q^*) + exp(\mu_{t_i} + \sigma_{t_i} q^*)}{2} exp(\mu_{t_i} + \frac{\sigma_{t_i}^2}{2})$$

where q^* is the $1 - \alpha/2$-quantile of the standard normal distribution. Here we take $\alpha = 0.1$.

As N_t can be measured by different aggregate variables, the prediction is made respectively. With the current status N_{t_i}, the future status $N_{t_{i+1}}$ is predicted as then:

$$N_{t_{i+1}} = [x_{t_i} N_{t_i}] \tag{22}$$

which is $x_{t_i} N_{t_i}$ rounded to the nearest integer. In this way, the evolution of ISA can be predicted.

5. EXPERIMENTAL VALIDATION

5.1 Set-Up

Dataset. The proposed Parameterized Social Activity Model (PSAM) is evaluated on 3 publicly-available datasets that represent different types of social activities: Facebook wall-post dataset [28], Facebook friend-request dataset [28], and Citeseer co-authorship dataset [15].

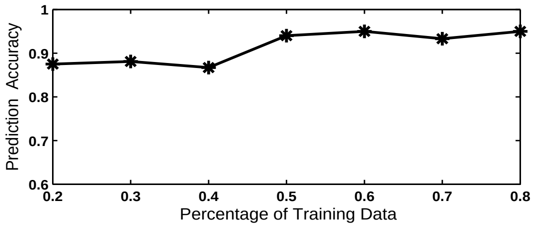


Figure 2: Accuracy of 90% Confidence interval of parameterized social activity model on Facebook friend-request dataset, when N(t)=|V(t)|

Table 2: Accuracy of 90% confidence interval of parameterized social activity model for 3 measures of social activity on Facebook friend-request dataset

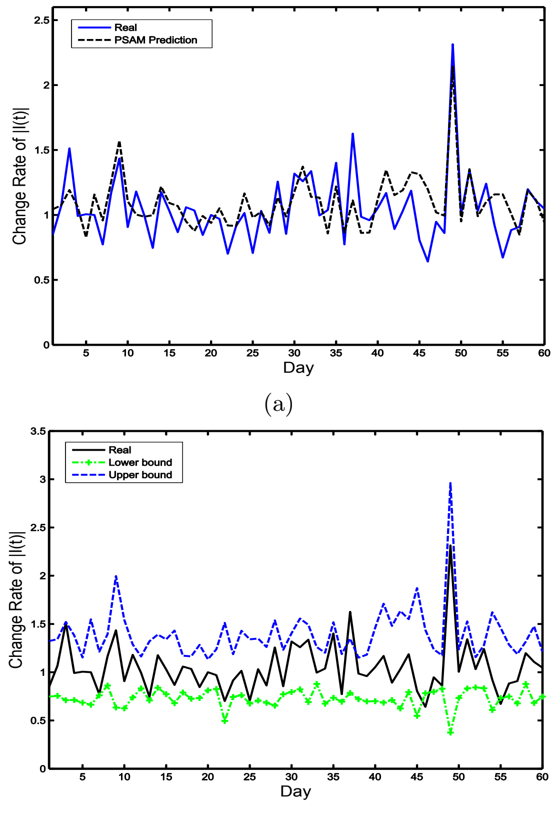
	N=|I(t)|	N=|E(t)|	N=|V(t)|
Accuracy of 90% CI	0.95	0.933	0.95

The Facebook datasets [28] are crawled from New Orleans regional network in Facebook. The friend-request dataset contains 905,565 records from 61,096 users over 850 days, while the wall-post dataset contains 876,993 posts between 46,952 users over 220 weeks (user attributes and textual content of wall posts are not available in the dataset). Each record is a tuple of $\langle UserID_1, UserID_2, Timestamp \rangle$. A record in the friend-request dataset denotes the time when $UserID_1$ adds $UserID_2$ as friend, while a record in the wall-post dataset indicates the time when $UserID_1$ posts on the wall of $UserID_2$. The Citeseer dataset contains 283,174 authors with 451,305 papers published from 1980 to 2005.

In the Facebook wall-post dataset, ISA is measured on a weekly basis, and activity features are extracted from them. The change rate is calculated with ISAs on every pairs of two successive weeks. In the Facebook friend-request dataset, to evaluate the predicting model on different time intervals, the interval of ISA is set to one day. For example, for the ground truth over 60 days, we construct 60 different ISAs and calculate a total of 60 change rates for them. In the Citeseer dataset, finally, the ISA and activity features are generated in a similar way. ISA is constructed with papers published every year and the corresponding authors. The change rate is then calculated with ISAs on every two successive years.

Feature selection. On each dataset, we extract 14 activity features as introduced in Table 1. After that, the *Principal Component Analysis (PCA)* is applied to eliminate the redundancy in the feature space, and to remove some noise. This also allows us to easily include more features into the framework without sacrificing performance. Then, top principal components with more than 80% of total data variation are used as the predicting vector.

Evaluation Metrics. To validate our method, we use the widely accepted evaluation metrics: Correlation and 90% Confidence Interval (CI). In particular, when we assume that the ground truth dataset is X and corresponding predicted dataset is Y, $\bar{x}$ and $\bar{y}$ denotes the mean values of X and Y respectively, and the correlation between X and Y is calcu-



(a)

(b)

Figure 3: Comparison of parameterized social activity model prediction and ground truth on Facebook friend-request dataset over 60 days, $N(t) = |I(t)|$. (a): PSAM prediction and ground truth, (b): upper and lower bound of PSAM 90% CI with ground truth

lated as follows:

$$Corr(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Baseline. Three baseline approaches are used for performance comparison with our PSAM model: linear regression (LR), cubic spline interpolation (CSI), and the average of ground truth of previous 5 ISAs (denoted as AVE).

5.2 Evaluation

In the experiments, we evaluate PSAM with one connective activity: friend-request on Facebook, and two interactive activities: wall-posting on Facebook and co-authorship on Citeseer. In particular, we evaluate PSAM on wall-posting data for the entire dataset as well as a subset of users. Meanwhile, we treat co-authorship as a dynamic and interactive activity – two scholars co-authoring a paper this year may or may not collaborate in next year. This is different from existing works on social network evolution, in which co-authorship links are "permanent" – once two scholars co-author a paper, they become co-authors permanently. Note that we are studying a different problem than network evolution – identifying active collaborations and predicting future collaborations. In this context, it is more reasonable to recognize the dynamics of co-authorship activities.

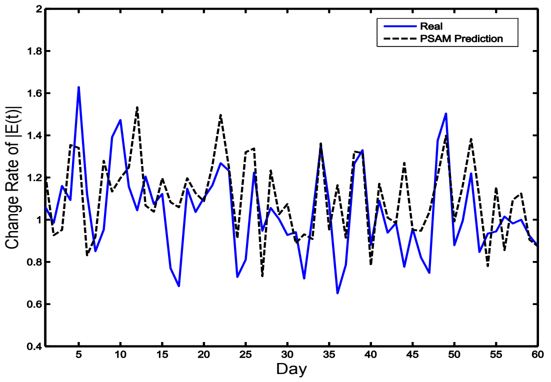


Figure 4: Comparison of parameterized social activity model prediction and ground truth on Facebook friend-request dataset, $N(t) = |E(t)|$

A. Facebook Friend-request Activity.

The Facebook dataset records daily friendship-requests for more than 2 years. As it is unnecessary to train and test the model for such a long period, we randomly select a subset of records on 300 successive days. In this experiment, we take three different measures of $N(t)$: $N(t) = |V(t)|$, $N(t) = |I(t)|$, and $N(t) = |E(t)|$. The parameterized evolution model is trained and tested against each measure.

The first group of experiments are conducted with $N(t) = |V(t)|$. To test the prediction accuracy, some of the records are randomly selected for training, and the rest for testing. We start by using 20% of the data for training (and 80% for testing), and gradually increase the portion of training data to 80% (hence, 20% for testing). Figure 2 shows the accuracy of 90%-CI with different percentages of training data (please note that the Y-axis starts at 0.6). Overall, the accuracy increases when we use more data for training. However, as shown in the figure, the PSAM approach has achieved 87.5% prediction accuracy even when only 20% of data (60 days) is used for training. The maximum accuracy is 95%, obtained when 80% of the data is used for training.

In the second group of experiments, the parameterized social activity model is compared with three baseline approaches on all measures of $N(t)$. With each measure, we compare our model prediction with the ground truth and measure the accuracy of 90% CI. Then the correlation of all methods with ground truth is evaluated.

First, when $N(t) = |I(t)|$, Figure 3(a) shows the comparison of PSAM prediction and ground truth on 60 random days. The model forecasts the up and down oscillation and predicts the change rate accurately, especially, for days 9, 28, 50 and 51. Figure 3(b) compares the upper and lower bounds of 90% CI with ground truth on the same days. A big burst of change rate 2.31 happens on day 49, i.e. an outlier. From Figure 3(a), we can see the model successfully predicts this burst. In Figure 3(b), the 90% CI covers this outlier without introducing a large variance. Among the 60 days, 57 data points falls into the 90% CI.

Overall the PSAM prediction is precise, though there are still a few points where predictions are not perfect. Considering the trade-off between the model usability and the

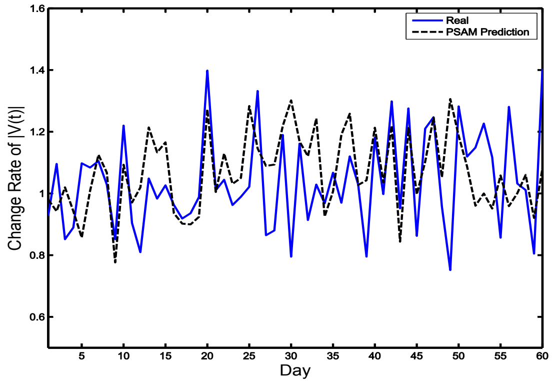


Figure 5: Comparison of parameterized social activity model prediction and ground truth on Facebook friend-request dataset, $N(t) = |V(t)|$

Table 3: Correlation of PSAM and 3 baseline methods against ground truth on Facebook friend-request dataset, when $N(t) = |V(t)|$, $|E(t)|$, and $|I(t)|$

Correlation	PSAM	LR	CSI	AVE
$N(t) = \|I(t)\|$	0.700	0.035	0.126	-0.671
$N(t) = \|E(t)\|$	0.515	0.059	0.222	-0.673
$N(t) = \|V(t)\|$	0.302	0.004	-0.026	-0.669

predicting accuracy, we suggest that 14 features are appropriate to ensure the simplicity as well as a high accuracy. For $N(t) = |E(t)|$ and $N(t) = |V(t)|$, the comparisons of PSAM prediction and ground truth on 60 random days are shown in Figure 4 and Figure 5 respectively. Comparing Figure 4 and Figure 5, we observe that the change rates of $|V(t)|$ and $|E(t)|$ evolve within the range $[0.75, 1.40]$ and $[0.65, 1.63]$ respectively, which are very close. The accuracy of 90% CI for the three social activity measures is shown in Table 2.

Finally, we compare PSAM with the baseline approaches. Table 3 illustrates the correlations of all methods with the ground truth, as $N(t)$ is measured by $I(t)$, $E(t)$, and $V(t)$. PSAM outperforms baselines with a high correlation. It has the highest correlation when predicting $I(t)$. From this group of experiments, it can be concluded that the PSAM makes effective and accurate prediction, even for outliers. The reason is that it is parameterized by the current activity features, which enables the model to incorporate the current network characteristics and therefore predict social activity accurately.

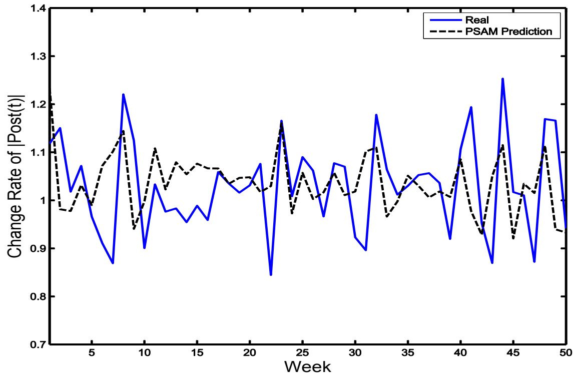


Figure 6: Comparison of parameterized social activity model prediction and ground truth on Facebook wall-post dataset over 50 weeks, $N = |Post(t)|$

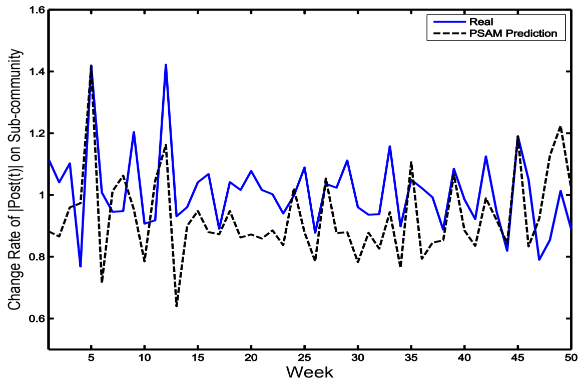


Figure 7: Comparison of PSAM prediction and ground truth on a 3000-user sub-community of Facebook wall-post data, $N = |Post(t)|$

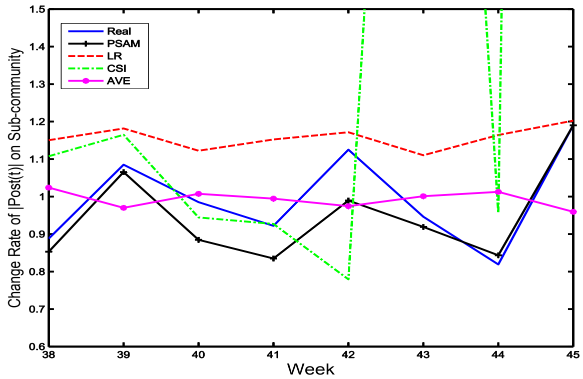


Figure 8: Details of PSAM prediction, baselines and ground truth from week 34 to week 42 in Figure 7

B. Facebook Wall-post Activity.

In the Facebook wall-post dataset, social activities are modeled as "posting on a friend's wall". We aim to predict the total number of posts in future, i.e. $N(t) = |Post(t)| = |I(t)|$. To demonstrate the effectiveness of the PSAM at different predicting intervals, the ISA interval on this dataset is set to be one week, instead of one day. We randomly select a period of 50 weeks for testing and use the rest 170 weeks for training. On this dataset, we evaluate the accuracy of PSAM, and compare it with three baselines.

Figure 6 shows the comparison of PSAM prediction and ground truth. The predictions generally match the up and down oscillation of ground truth. Some data points are accurately predicted, while others are not perfect yet. For the 50 weeks, the accuracy of 90% CI is 0.8, as shown in Table 4. Furthermore, to validate the performance of PSAM on arbitrary set of users, we choose 3,000 most active users from the wall-post dataset and try to predict the activities among them. In Figure 7, the PSAM prediction is compared with

Table 4: Accuracy of 90% confidence interval of PSAM on overall Facebook wall-post dataset and a subset of 3000 users

	Wall-post Overall	Wall-post Subset
Accuracy of 90% CI	0.8	0.94

Table 5: Correlation of PSAM and 3 baseline methods against ground truth on overall Facebook wall-post dataset and subset of 3000 users

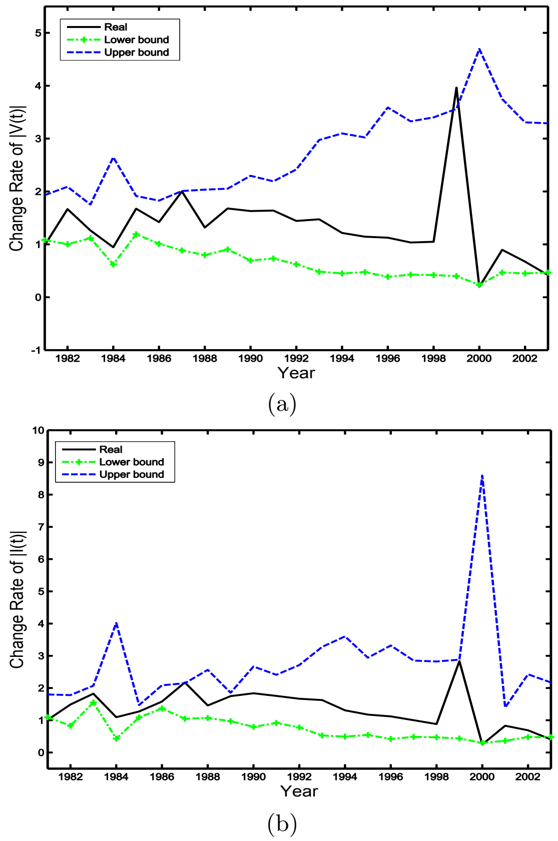
Correlation	PSAM	LR	CSI	AVE
Overall	0.215	0.194	0.045	-0.197
Subset	0.481	0.291	0.138	-0.228

ground truth. We observe that the prediction accuracy is even better than that on the entire dataset. The accuracy of 90% CI is 0.94, as shown in Table 4.

Next, PSAM is evaluated against baselines on both overall wall-post dataset and the sub-community. To illustrate the performance of PSAM and baselines in detail, a section from week 38 to week 45 in the wall-post sub-community is enlarged in Figure 8. In this figure, we observe that the plot of LR is very flat and completely above the ground truth. For CSI, the predictions are far away from ground truth on some points, so they are not plotted in the figure. Predictions of AVE do not reflect ground truth on almost all points. Compared to baselines, predictions of PSAM are much more accurate. Table 5 shows the correlation of all methods against ground truth. Overall AVE is the worst among all methods and PSAM is the best. The correlation of PSAM on the subset is much higher than that on the overall dataset. It implies that PSAM may perform better on mid size communities.

C. Citeseer Co-authorship Activity.

As the Citeseer dataset only covers 25 years, relatively short for the time interval of ISA, overall dataset is used for both training and testing. Due to the page limit, we only show partial results of two measures: $N(t) = |I(t)|$ and $N(t) = |V(t)|$. Figure 9(a) shows the upper and lower bounds of 90% CI and ground truth on each year when $N(t) = |V(t)|$. In Figure 9(b), they are compared for $N(t) = |I(t)|$. Although the data set is small and the time interval of ISA is large, the 90% CI still has an accuracy of 0.82 for both measures. The prediction presents a good coverage of the ground truth. In all, the 90% CI of PSAM has more than 0.8 accuracy on all three datasets, which indicates that the model is effective to simulate the ISA evolution and applicable to predict various social activities. At the same time, PSAM is validated with different time intervals of ISA. Besides that, different measures of $N(t)$ also indicate that PSAM can make accurate prediction on multiple aspects and granularities.



(a)

(b)

Figure 9: 90% CI of PSAM and ground truth on overall Citeseer data (a):$N(t) = |V(t)|$ (b) $N(t) = |I(t)|$

6. RELATED WORK

Social network evolution has been explored from different perspectives in recent years. They consider social networks formed in different environment: world-wide web, blogger networks, online friendship networks, academic co-authorship, etc. Existing works can be roughly categorized into three groups: static network mining, microscopic evolution prediction, and time-evolving structure analysis.

Some structural properties are discovered by mining the snapshots of the static network. The well-known small-world phenomenon is observed in [31]. A study on the web shows that its average diameter is small and the web forms a small-world network [2]. These studies revealed important properties but they were performed only on the static graphs.

The prediction of microscopic evolution pays more attention to the addition of new edges and new vertices [7]. The classic E-R model simulates the network growth when the edges between vertices are added randomly [12]. Different from that, preferential attachment of new vertices is proposed to capture the power-law degree distribution [5][11]. Another microscopic evolution model is proposed with nodes arriving at a pre-specified rate and selecting their lifetimes [20]. In [8], membership vectors of current members are used to address the co-evolution of the network topology and membership. Later, topological features and relations are explored to predict when a new link will appear [27].

More recent studies explore the state transition of individuals in a social network with a space of finite states. In [6], a model called NLDS, is presented to model the viral propagation and find out the epidemic threshold condition. Subsequently, to predict the popularity of news, a model based on the web site design is proposed to simulate voters' behavior [19]. These two models are only applicable to evolutions with countable future states. Different from them, the social activity evolution in our model has an infinite future state space. Therefore, our model can handel more situations on a broader perspective.

Time-evolving structure analysis focuses on the structural features and their evolution patterns [3]. The forest fire model [21] explains the densification and shrinking diameters over time. Different behavior scaling in degree distribution is analyzed on various online social networks [1] and the co-evolution of social and affiliation networks is addressed [33]. By analyzing a co-authorship network and a phone-call network, Palla et al. revealed some activity patterns and the influence of geographic locations [25]. In [23], researchers track the social events by studying the interplay between textual topics and network structures.

In these studies, the influence of structural properties is examined at the individual level and the measures of evolution are based on the cumulative additions of new vertices.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we study and predict aggregate social activity with various measures. By observing that social connections are insufficient to determine a wide variety of user behaviors, we propose a continuous-time stochastic process to model the social activity evolution. The model is parameterized by activity features and integrates Wiener Process as a component, in order to address the randomness of the process and ensure dynamics therein. Experiments on three real datasets of different social activities reveal that our proposed parameterized social activity model (PSAM) can predict aggregate social activity accurately and outperforms other competing methods.

Many research issues are remaining for future work. We first plan to investigate other types of social activities in modeling and predicting future states of a social network. Second, while the accuracy of our proposed model is promising and superior to three baseline approaches, we will explore other ideas to improve the accuracy further.

8. ACKNOWLEDGEMENT

Both Shu Huang and Dongwon Lee were in part supported by NSF DUE-0817376 and DUE-0937891 awards. Bo Luo was partially supported by NSF OIA-1028098 award.

9. REFERENCES

[1] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW*, 2007.

[2] R. Albert, H. Jeong, and A.-L. Barabasi. Diameter of the world-wide web. *Nature*, 401:130–131, 1999.

[3] S. Asur, S. Parthasarathy, and D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. In *KDD*, 2007.

[4] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD*, 2006.

[5] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, October 1999.

[6] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *ACM TISSEC.*, 10, January 2008.

[7] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD*, 2011.

[8] Y.-S. Cho, G. Steeg, and A. Galstyan. Co-evolution of selection and influence in social networks. In *AAAI*, 2011.

[9] D. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *PNAS*, 2010.

[10] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg. Echoes of power: language effects and power differences in social interaction. In *WWW*, 2012.

[11] E. Elmacioglu and D. Lee. Modeling idiosyncratic properties of collaboration networks revisited. *Scientometrics*, 80(1):195–216, July 2009.

[12] P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

[13] H. Fei, R. Jiang, Y. Yang, B. Luo, and J. Huan. Content based social behavior prediction: a multi-task learning approach. In *CIKM*, 2011.

[14] J. Huang, H. Sun, J. Han, H. Deng, Y. Sun, and Y. Liu. Shrink: a structural clustering algorithm for detecting hierarchical communities in networks. In *CIKM*, 2010.

[15] J. Huang, Z. Zhuang, J. Li, and C. L. Giles. Collaboration over time: characterizing and modeling network evolution. In *WSDM*, 2008.

[16] I. Karatzas and S. E. Shreve. Brownian motion and stochastic calculus. page 149. Springer, Aug. 1991.

[17] G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):80–90, 2006.

[18] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Link Mining: Models, Algorithms, and Applications*, pages 337–357. Springer New York, 2010.

[19] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *WWW*, 2010.

[20] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD*, 2008.

[21] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD*, 2005.

[22] C. W.-k. Leung, E.-P. Lim, D. Lo, and J. Weng. Mining interesting link formation rules in social networks. In *CIKM*, 2010.

[23] C. X. Lin, B. Zhao, Q. Mei, and J. Han. Pet: a statistical model for popular events tracking in social communities. In *KDD*, 2010.

[24] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, and J. Chayes. We know who you followed last summer: inferring social link creation times in twitter. In *WWW*, 2011.

[25] G. Palla, A.-L. Barabasi, and T. vicsek. Quantitative social group dynamics on a large scale. *Nature*, 2007.

[26] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz. Modeling and predicting behavioral dynamics on the web. In *WWW*, 2012.

[27] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla. When will it happen?: relationship prediction in heterogeneous information networks. In *WSDM*, 2012.

[28] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in facebook. In *WOSN*, 2009.

[29] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *KDD*, 2011.

[30] X. Wang, L. Tang, H. Gao, and H. Liu. Discovering overlapping groups in social media. In *ICDM*, 2010.

[31] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[32] S.-H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, and H. Zha. Like like alike: joint friendship and interest propagation in social networks. In *WWW*, 2011.

[33] E. Zheleva, H. Sharara, and L. Getoor. Co-evolution of social and affiliation networks. In *KDD*, 2009.