The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

# Dynamic Detection of Communities and Their Evolutions in Temporal Social Networks

**Yaowei Huang,* Jinghuan Shang,* Bill Y. Lin,* Luoyi Fu, Xinbing Wang**
{14330222150355,elicassion,yuchenlin,yiluofu,xwang8}@sjtu.edu.cn
Department of Computer Science and Engineering
Shanghai Jiao Tong University
800 Dongchuan Road, Shanghai, China 200240

## Abstract

In this paper, we propose a novel community detection model, which explores the dynamic community evolutions in temporal social networks by modeling temporal affiliation strength between users and communities. Instead of transforming dynamic networks into static networks, our model utilizes normal distribution to estimate the change of affiliation strength more concisely and comprehensively. Extensive quantitative and qualitative evaluation on large social network datasets show that our model achieves improvements in terms of prediction accuracy and reveals distinctive insight about evolutions of temporal social networks.

## Introduction

A community in social networks is a cluster of nodes with more intense interactions amongst its members than others. Detecting communities in temporal social networks and studying their evolutions are very beneficial but challenging. In this paper, we study community detection and the evolution of communities in temporal social networks, by modeling temporal strength between users and communities.

Figure 1 shows the overview of our model. The input is an interaction network among users, where the directed edges have two time-stamps for two interacting users. The output is temporal distribution of the affiliation strength of different communities for each user. Our main contributions in this paper are as follows:

1. *Novel Perspective*: To the best of our knowledge, we are among the first to study community detection and evolution by modeling temporal strength between users and communities and observe flows of the membership strength of users among multiple communities as *community evolution*.
2. *Novel Model*: We model the continuous relationship between users and communities with normal distribution.
3. *Better Performance:* Extensive experiments show our model outperforms strong baseline methods by substantial margins. We also present real-world applications.

*The first three authors contributed equally.



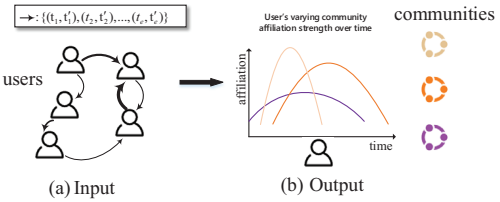



Figure 1: (a) is the input temporal social network. (b) is the output distributions.

## Problem Formulation

**Temporal Interaction Network** A temporal interaction network is defined as a directed unweighted graph $G = (V, E)$. Each edge $e \in E$ is associated with two time stamps $t_1, t_2$. The time stamps represents the posting time ($t_1$) and mentioning time ($t_2$). A directed edge $(u, t_1, v, t_2)$ indicates the communication of users at different moments. [1]

**Affiliation Strength** A nonnegative parameter $F_{uc}$ represents the *affiliation strength* of a user $u$ in a community $c$. $F_{uc} = 0$ means node $u$ is not affiliated to community $c$. Thus, we use $P_{uc}(t)$ to represent the weights of $F_{uc}$ at time $t$. Finally, we have $\pi_{uc}(t) = P_{uc}(t)F_{uc}$ as the temporal strength between a user $u$ in a community $c$ at time $t$.

**Probability of User Interaction** We denote $p_{u,v,t_1,t_2}$ as the probability of the existence of an edge $(u, t_1, v, t_2)$ in a *Temporal Interaction Network*. We assume the connection between users are through all internal communities with different contributions. The probability of an interaction between two users through a particular community $c$ is $\pi_{uc}(t_1)\pi_{vc}(t_2)$. $p(u, t_1, v, t_2)$ is calculated in the following equations, where $\mathcal{C}$ is the set of all communities.

$$\begin{aligned} H &= \sum_{c \in \mathcal{C}} \pi_{uc}(t_1)\pi_{vc}(t_2), \\ p(u, v, t_1, t_2) &= 1 - \exp(-H), \end{aligned} \tag{1}$$

The problem we aim to solve is how to better model the such probability of edges in temporal social networks.

[1] This can be derived from the various types of interactions such as citing papers, re-tweeting and commenting on social media.

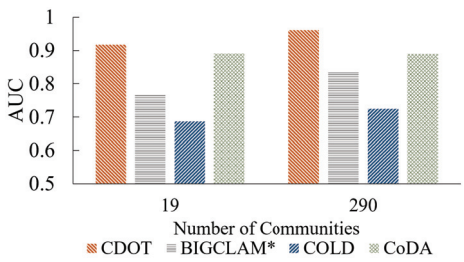



Figure 2: Link Prediction Performance

## Modeling

We assume a network as a result generated by a variant of the community-affiliation graph model. Given a temporal social network, our model produces a bipartite graph where the nodes on one side represent the *users* in the social network $G$ and the nodes on the other side represent *communities*. We assume $P_{uc}(t)$ obeys *normal distribution*, which means $P_{uc}(t) = \frac{1}{\sqrt{2\pi}\sigma_{uc}} \exp \frac{-(t-\mu_{uc})^2}{2\sigma_{uc}^2}$. Thus, we have $\mu_{uc}$ and $\sigma_{uc}$ as the mean and the variance values of the normal distribution with respect to $\pi_{uc}(t)$. Our goal of learning is to maximize the log-likelihood, so that $\hat{F}, \hat{\mu}, \hat{\sigma} = \arg\max_{F \geq 0, \sigma > 0} l(F, \mu, \sigma)$ where

$$l(F, \mu, \sigma) = \sum_{(u,v,t_1,t_2) \in E} \log(1 - \exp(-H)) - \sum_{(u,v,t_1,t_2) \notin E} H. \quad (2)$$

In our supplementary material, we talk about the advantages of using normal distribution and how we sample the negative edges to improve the time complexity.

## Evaluation

### Evaluation Setup

**Datasets** Based on MAG (Microsoft Academic Graph), we create two datasets (M200, BD) for our quantitative evaluation and qualitative evaluation respectively[2]. Another dataset is used for qualitative evaluation consists of papers under the research topic *Big Data*, containing 81K nodes, 2M edges and 120K papers with 25 communities.

**Baseline Methods** We compare our model (CDOT) with several following state-of-the-art competitors, namely BIGCLAM (**?**), CoDA (**?**) and COLD (**?**)[3].

### Quantitative Evaluation

**Community Detection (Link Prediction)** Figure 2 shows the AUC scores of the four models. Our model demonstrates better performance on link prediction task than other models. The result reveals that our model is able to capture dynamic strength between users and communities.

[2] Based on these papers, we gain their authors and publishing years. M200 contains 318K nodes, 4M edges and 500K papers with two scales of the number of communities, 19 and 290.

[3] Our model, CoDA and BIGCLAM take only nodes and edges as input, while COLD additionally utilizes titles of papers.

| Model | 19 Communities | 290 Communities |
|---|---|---|
| CDOT | **2.734** | 2.363 |
| COLD | 4.693 | **2.161** |

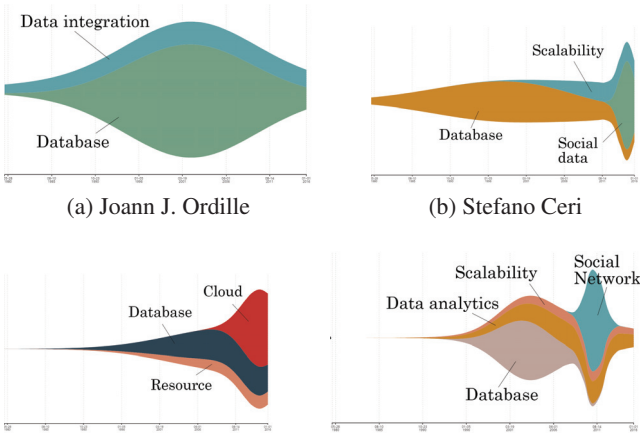
Table 1: Nlog Measurement Result

**Temporal Modeling (Time Stamp Prediction)** Time stamp prediction is to estimate the occurring time of a previously unseen document.

We take the average value of all of them from input edges as the result of nlog measurement. [4] Table 1 illustrates the scores of nlog measurement.

### Qualitative Evaluation – Application

Utilizing the temporal social network, our model is able to detect user strength of affiliation among truth or latent communities. Furthermore, we can demonstrate the membership of users to communities at a time to grasp the whole picture of communities. An example is shown in Figure 3.




(a) Joann J. Ordille (b) Stefano Ceri

(c) Danilo Ardagna (d) Hakan Ferhatosmanoglu

Figure 3: Temporal activation of researchers among fields. Names are listed below the corresponding pictures. The horizontal axis represents the time, from 1980 to 2016. The vertical one suggests the affiliation strength. Each block(color) is on behalf of a community, or a field of study equally. The overall wavy shape reflects the variation of affiliation strength through one's research career.

[4] Note that BIGCLAM and CoDA do not support the temporal prediction.