A Longitudinal Social Network Clustering Method Based on Tie Strength

Zhiyong Zhang
Department of Psychology
University of Notre Dame
Notre Dame, IN, USA
zzhang4@nd.edu

Mao Ye
Department of Statistics
Purdue University
West Lafayette, IN, USA
ye207@purdue.edu

Yijie Huang
Quantitative Methods in Social Science
Columbia University
New York, NY, USA

Nan Sun
Department of Computer Science
Sichuan University
Sichuan, China

***Abstract*—Longitudinal social network clustering is an emerging research area with many applications. Previous research typically focuses on the development of the clusters in the longitudinal network. In this paper, we propose an alternative method for longitudinal social network clustering, in which we assume that the clustering and the evolution of the network are the results of its inner structure, the strength of the ties among the nodes in the network. We estimate the strength of the ties based on the evolution of the network over time through a continuous Markov process and then clustering the network based on the strength of the ties of the whole network. A simulation study shows that the proposed method performs well under a variety of conditions. The application of the method is illustrated through the analysis of a real set of data.**

***Keywords*-Longitudinal network; clustering; stochastic actor-oriented model; Bayesian method**

I. Introduction

A social network is a systematic depiction of the living context of individuals. Focusing on the relationship among social entities, social network analysis is widely used in the social and behavioral sciences. A network can be characterized as a set of nodes joined in pairs by edges or ties [1]. Specifically, each node in the network represents an individual, also called an actor; and a directed or undirected tie between two nodes describes the relationship between them in reality. Typically, a binary variable X_{ij} is used to denote the relationship, with 1 indicating the existence of a tie from node i to node j, and 0 not. A complete social network can be represented as an adjacency matrix $\mathbf{X} = (X_{ij}), i, j = 1, \ldots, g$ with g being the total number of nodes in the network. If a social network evolves throughout time, a longitudinal social network is observed and can be represented by a series of adjacency matrix $\mathbf{X}(t_l) = (X_{ij}(t_l))$ for $l = 1, ..., M$, where M is the total number of times that the network is observed and t_l denotes the time when the network is observed.

Network clustering, or community detection, is an active research field in network analysis [2]. Clustering can be conducted for both undirected and directed social networks. Directed social networks are more common in the social and behavioral sciences and they usually contain more information. For example, in studying a friendship network, only a directed network can reflect the orientation of relationships.

Two popular approaches have been proposed to cluster the directed networks. One approach is to extend the existing methods for undirected networks to the directed ones such as the methods based on modularity [3], cut-based measures [4], and spectral clustering [5]. Another approach to clustering directed networks is through transformation. The typical method is to cluster different communities via a two-step procedure [6]. In the first step, a directed network is "symmetrized", and in the second step, the resulting symmetrized matrix is clustered by selected undirected clustering algorithms. By symmetrizing a network, we can transform the directed network into an undirected and weighted network.

While static network clustering (directed or undirected) has already been studied extensively [7], [8], dynamic network clustering received relatively less attention. Previous methods generally conduct dynamic network clustering using a two-step procedure: (1) identifying the clusters based on the static snapshots of the time-evolving network, and (2) detecting the change points throughout the time according to the different partitions over time [8].

Other dynamic network clustering approaches are also available, which only differ in specific methods of clustering a static snapshot and detecting the change point in the two-step procedure. The existing methods have focused on networks in which ties are regarded as brief events such as email and cell-phone communications among actors. However, in many social networks on friendship, trust, and cooperation, a tie or relationship often builds gradually and endures over time as opposed to being created and terminated spontaneously. For such networks, the two-step

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

procedure may not be appropriate for at least two reasons. First, the current states of a network cannot be regarded as a snapshot captured from its evolution. A network observed at the current time depends on previous networks and can also predict future networks. Simply dividing the dynamic network evolution process into discrete parts without considering its inner causality may ignore important evolution information. Second, the existing methods may not reflect the tendency of components forming different communities at the macroscopical level, and do not provide informative and intuitive indications on the inside structure of a network extracted from the observed changes.

Therefore, the purpose of this study is to propose a new dynamic network clustering method to better cluster longitudinal networks by integrating the information of dynamic transition in the structure of social networks. In the rest of the paper, we first present our proposed method. Then, through a simulation study, we evaluate factors that affect the clustering performance. After that, we show its application by clustering a real longitudinal social network. Finally, we discuss future directions and limitations of the study.

II. Longitudinal Social Network Clustering

For a longitudinal friendship network, the strength of the friendship, or ties, is crucial and, therefore, we propose to cluster based on such strength of ties. The strength of the ties measures the intensity and tendency of the connection between two actors. It is the comprehensive result of various unknown latent variables or factors [9]. In the literature, a number of models based on the distance measure have been proposed to measure the strength of the ties considering the latent variables or the topology of a social network [10]–[16]. The general idea of this paper is to evaluate the strength of the ties according to how a network evolves throughout time. To measure the strength of a tie of two actors, we assume that the tie follows a continuous Markov process. The strength of this tie can then be measured by the parameters, or a function of them, of the probability distributions of the Markov process. Using the estimated strength of ties, we can form an undirected and weighted adjacency matrix. Then, a spectral analysis algorithm [17], [18] based on k-medoids algorithm [19] can be used to cluster the longitudinal network based on the undirected and weighted adjacency matrix.

In the rest of this section, we will first discuss the basic assumptions of our method and how we evaluate the strength of the ties based on the probability distribution of the Markov process. After that, we will propose a Bayesian parameter estimation method of the strength and illustrate its use through a data experiment. Then, we will briefly introduce the spectral clustering algorithm.

A. Assumptions

Let $X_{ij}(t) = 1$ denote the existence of a tie between actors i and j at time t, and $X_{ij}(t) = 0$ not. To estimate the strength of a tie, we assume that the tie follows a weakly stationary continuous Markov process with the rate λ_0 switching from 0 to 1, e.g., forming a tie from one time to the next time, and rate λ_1 for switching from 1 to 0, e.g., breaking a tie from one time to the next time. Both $X_{ij}(t)$ and $X_{ji}(t)$ are assumed to be independent and identically distributed stochastic processes. Our strength measure focuses on the evolution of a tie throughout the time. Let $P_{lk}(\Delta t)$ be the probability for the Markov process transferring from state l to state k ($l = 0, 1, k = 0, 1$), over a time interval Δt, that is, the Markov process begins at state l at certain time point t_0 and ends at state k at time $t_0+\Delta t$. According to [20], solving the Kolmogorov forward equation

$$\begin{aligned}\frac{d}{d\Delta t}P_{00}(\Delta t) &= \lambda_1 P_{01}(\Delta t) - \lambda_0 P_{00}(\Delta t)\\ &= -(\lambda_0+\lambda_1)P_{00}(\Delta t) + \lambda_1\end{aligned}$$

leads to the transition probability

$$P_{00}(\Delta t) = \frac{\lambda_1}{\lambda_0+\lambda_1} + \frac{\lambda_0}{\lambda_0+\lambda_1}e^{-(\lambda_0+\lambda_1)\Delta t}.$$

Similarly, we have

$$P_{11}(\Delta t) = \frac{\lambda_0}{\lambda_0+\lambda_1} + \frac{\lambda_1}{\lambda_0+\lambda_1}e^{-(\lambda_0+\lambda_1)\Delta t}.$$

Given $P_{01}(t) = 1 - P_{00}(t)$, we have:

$$\lambda_0 = -\frac{P_{01}(\Delta t)\times \ln(P_{11}(\Delta t) - P_{01}(\Delta t))}{\Delta t(1 + P_{01}(\Delta t) - P_{11}(\Delta t))}, \quad (1)$$

$$\lambda_1 = -\frac{(1 - P_{11}(\Delta t))\times \ln(P_{11}(\Delta t) - P_{01}(\Delta t))}{\Delta t(1 + P_{01}(\Delta t) - P_{11}(\Delta t))}. \quad (2)$$

Because the Markov process is irreducible and aperiodic, we have

$$\theta_0 = \lim_{\Delta t\to\infty} P_{00}(\Delta t) = \frac{\lambda_1}{\lambda_0+\lambda_1},$$

which can be interpreted as the portion of the time that a process stays at a state in the long run. The quantity $\theta_1 = \lim_{\Delta t\to\infty} P_{11}(\Delta t) = \frac{\lambda_0}{\lambda_1+\lambda_0}$ can be used to measure the strength of the tie between two actors in a social network considering not only the transition rate but also the probability of the stationary distribution. The higher its value is, the stronger the tie is.

For illustration, consider two longitudinal social networks with two actors with the information of the ties shown below:

$$\text{Network}_1 = [0,0,0,0,0,0,1,1,1,1,1,1],$$

$$\text{Network}_2 = [0,1,0,1,0,1,0,1,0,1,0,1],$$

with 1 representing a tie between the two actors. There are 12 waves or times of data in both networks. In the first network, a tie is not observed in the first six time points but observed in the last six time points. On the other hand, the tie in the second network builds and breaks alternatively. Although the total number of ties observed in the two networks are the same, they display distinct patterns. If the tie represents the friendship of two actors, it is clearly not stable in the second network. The first network shows a stable friendship after building up. The two individuals in the network may not have enough interaction to form a friendship with each other in the first six time points. However, once they build the friendship, they maintain it. Intuitively, we should believe that the tie in the first network is stronger than that in the second network. The statistic, $\frac{\lambda_1}{\lambda_0+\lambda_1}$, provides a good measure of the strength in distinguishing the different patterns as we will show later.

Note that the way a tie strength is defined values the continuation and duration of the longitudinal dyadic relationship. It works best for friendship network when identifying the common cause underlying a longitudinal network. If the purpose of the network analysis changes, the current tie strength estimation might not be appropriate anymore.

B. A Bayesian Estimator

Bayesian estimation methods can be used to estimate $\lambda_0/(\lambda_1 + \lambda_0)$ from data collected for a network. For an observed longitudinal network, we have data $X_{ij}(t_1), X_{ij}(t_2), ..., X_{ij}(t_M)$ and $X_{ji}(t_1), X_{ji}(t_2), ..., X_{ji}(t_M)$. The time intervals between two consecutive observations is $\triangle t = t_{l+1} - t_l,\ \forall l \in \{1, 2, ..., M-1\}$. Define

$$N_{hk(ij)} = \#\left\{(i,j) \mid X_{ij}(t_l) = h, X_{ij}(t_{l+1}) = k\right\},$$
$$h, k \in \{0,1\}\ \ l \in \{1, 2, ..., M-1\}$$

and

$$N_{hk(ji)} = \#\left\{(j,i) \mid X_{ji}(t_l) = h, X_{ji}(t_{l+1}) = k\right\},$$
$$h, k \in \{0,1\}\ \ l \in \{1, 2, ..., M-1\}.$$

For example, $N_{01(i,j)}$ is the number of the 0-to-1 transitions of ties from i to j. Since

$$N_{hk} = N_{hk(ij)} + N_{hk(ji)},\ h, k \in \{0,1\},$$

we have:

$$N_{01} \sim B(N_{01} + N_{00},\ P_{01}(\triangle t)),$$
$$N_{11} \sim B(N_{11} + N_{10},\ P_{11}(\triangle t)),$$

where B represents a binomial distribution. From Equations (1) and (2), we have

$$\frac{\lambda_0}{\lambda_1 + \lambda_0} = \frac{P_{01}(\Delta t)}{1 - P_{11}(\Delta t) + P_{01}(\Delta t)}.$$

To simplify the notation, let $p = P_{01}(\triangle t)$, $n = N_{00} + N_{01}$ and $k = N_{01}$. We denote the prior distribution and the posterior distribution of p as $\pi(p)$ and $\pi(p \mid n, k)$ respectively. Any informative or uninformative prior distribution can be used here. However, the same prior should be used in the estimation of the tie strength for all pairs of actors. To reduce the influence of priors, Jeffreys prior [21] is chosen for p in this study, that is $\pi(p) \sim Beta(\frac{1}{2}, \frac{1}{2})$. Using Bayes Theorem, the posterior is

$$\pi(p \mid n, k) \propto \pi(p)\pi(n, k \mid p) \propto \frac{\Gamma(1)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)} p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}$$
$$\times \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k},$$

which is also a Beta distribution $Beta(k + \frac{1}{2}, n - k + \frac{1}{2})$.

With the posterior distribution, an estimate of P_{01} through the posterior mean is

$$\hat{P}_{01} = \hat{p} = \int_{p \subseteq (0,1)} p \frac{\Gamma(n+1)}{\Gamma\left(k+\frac{1}{2}\right)\Gamma\left(n-k+\frac{1}{2}\right)} p^{k-\frac{1}{2}} \times$$
$$(1-p)^{n+k-\frac{1}{2}} dp = \frac{\frac{1}{2} + k}{1 + n} = \frac{\frac{1}{2} + N_{01}}{1 + N_{00} + N_{01}}.$$

Similarly, we have:

$$\hat{P}_{11} = \frac{\frac{1}{2} + N_{11}}{1 + N_{10} + N_{11}}$$

Then we have

$$\widehat{\frac{\lambda_0}{\lambda_0 + \lambda_1}} = \frac{\hat{P}_{01}}{1 - \hat{P}_{11} + \hat{P}_{01}} = \frac{\frac{\frac{1}{2}+N_{01}}{1+N_{00}+N_{01}}}{1 - \frac{\frac{1}{2}+N_{01}}{1+N_{00}+N_{01}} + \frac{\frac{1}{2}+N_{11}}{1+N_{10}+N_{11}}}.$$

This Bayesian estimation has two important advantages. First, it is robust and easy to calculate without numeric problem that might happen if maximum likelihood estimation is used. Second, it can handle missing data in network by simply removing the missing values.

C. Illustration of the Bayesian Estimation

We illustrate the Bayesian estimation of the strength of a tie through some simple data examples shown in Table I. Suppose that the tie represents the friendship between two individuals. Then, the data such as 1,1,1,1,0,0,0,0 in the table indicate that two individuals are friends from time 1 to time 4 but are not friends any more from time 5 to time 8. Using the Bayesian method, the strength of the friendship is 0.2941. A naive measure could be the simple ratio of ties of all time points, which is 0.5 in this case.

Table I shows three types of relationship between two individuals. For the first type of data, the Bayesian method can better distinguish the difference in the relationship while the ratios are exactly the same. The second type of data shows a remarkable change (from 0 to 1 or from 1 to 0) at the last one or two waves. Our method values

Table I
DATA EXPERIMENT

Type	Data	Ratio	Bayesian
1	1,1,1,1,0,0,0,0	0.5000	0.2941
	0,0,0,0,1,1,1,1	0.5000	0.7059
	0,1,0,1,0,1,0,1	0.5000	0.5070
	1,0,1,0,1,0,1,0	0.5000	0.4930
2	1,1,1,1,1,0	0.8333	0.6667
	1,1,1,1,1,0,0	0.7143	0.5000
	0,0,0,0,0,1	0.1667	0.3333
	0,0,0,0,0,1,1	0.2857	0.5000
3	0,0,0	0	0.2500
	0,0,0,0,0,0	0	0.1429
	1,1,1	1	0.7500
	1,1,1,1,1,1	1	0.8571

the current data more than the previous data because it is assumed the current relationship will influence future relationship more. Compared with the data 0,0,0,0,0,1, the data 0,0,0,0,0,1,1 provide more information about the tie between these two actors. Therefore, the estimated strength is higher for 0,0,0,0,0,1,1. Furthermore, comparing the data $1, 1, 1, 1, 1, 0, 0$ with the data $0, 0, 0, 0, 0, 1, 1$, the ratio of ties of the first set of data is much larger than that of the second data set. However, the data $1, 1, 1, 1, 1, 0, 0$ indicate that the tie breaks and, the data $[0, 0, 0, 0, 0, 1, 1]$ indicate that the tie forms during the last two waves. Therefore, our method estimates the same strength of the ties. The third type of data shows our method is less influenced by extreme data than the use of ratios.

D. Spectral Clustering

Given a longitudinal social network, the strength of the relationship between any two actors i and j can be estimated using our Bayesian method. Then, we can form an undirected strength matrix with diagonals being 0 to represent the strength of friendship among all actors. With the strength matrix, the existing methods for valued network clustering can be used to detect clusters among the actors. In this study, the spectral clustering is used in two steps. First, we transform the undirected strength matrix into a set of points in a distance space, whose coordinates are elements of eigenvectors. Second, we cluster the points via standard techniques such as k-medoids clustering.

Given the strength matrix $\mathbf{W}$ with g actors, where $w_{ij} = w_{ji} \geq 0, i = 1, \ldots, g, j = 1, \ldots, g$. The degree of an actor d_i is defined as $d_i = \sum_{j=1}^{g} w_{ij}$. The degree matrix $\mathbf{D}$ is defined as a diagonal matrix with the degrees $d_1, ..., d_g$ on the diagonal. With the information, the following algorithm can be used to conduct the spectral clustering.

1) Compute the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$.
2) Compute the m eigenvectors $u_1, ..., u_m$ corresponding to m smallest eigenvalues of $\mathbf{L}$.
3) Define $\mathbf{U} \subseteq \mathbb{R}^{g \times m}$ be the matrix containing $u_1, \ldots, u_m$ as columns.
4) Let y_i be the vector corresponding to the i-th row of $\mathbf{U}$, which is the coordinate of the actor i in the distance space.
5) Cluster the points with coordinates $y_i, i = 1, \ldots, g$ in $\mathbb{R}^m$ into k clusters based on k-medoids algorithm.

III. SIMULATION STUDY

To evaluate the performance of our method, we conduct a simulation study based on the actor-based model [22], [23]. The actor-based model is chosen for the purpose of generating longitudinal networks because of their popularity in the social and behavioral literature. However, our model should work for longitudinal networks generated by other models such as the actor-level models. The actor-based model was chosen for illustration. In the section, we first introduce the actor-based model and then present our simulation study.

A. Actor-based model

The actor-based model integrates several micro-mechanisms of a network [23]. In a stochastic actor-based model, a tie between two existing actors is typically regarded as, but not restricted to, the friendship between them, directed from one to another. A tie from i to j is interpreted as the actor i views j as a friend. The actors, in principle, control their outgoing ties but are also subject to constraints. At a given moment stochastically determined in the model, only one actor is allowed to change one tie.

According to [22], the stochastic actor-based model is based on several fundamental assumptions. First, the time parameter t is a continuous variable, suggesting that although observations are made at discrete time points, the development of a social network is the result of a continuous-time process. Second, the changing network is regarded as the outcome of a Markov process. Since the network ties are defined as states (like friendship) that tend to endure over time [22], the current states of the network fully mediate its future development in any given time. Third, the outgoing ties of each actor in the network are under her/his control. However, this does not mean that actors are able to change their ties at will, but are subject to the network structure. Fourth, at a given moment, only one selected actor has the chance to change one outgoing tie.

The objective function is a critical part of the actor-based model [22]. The objective function is defined as a linear combination of the components of the network as in,

$$f_i(\beta, x) = \sum_k \beta_k s_{ik}(x),$$

where $f_i(\beta, x)$ is the objective function for actor i that depends on the state x of the network. The function $s_{ik}(x)$ represents a component, or the so-called effect, of the network, indicating a specific tendency of change in the network. The parameter β_k is the weight of each effect $s_{ik}(x)$. If $\beta_k = 0$, it indicates that the corresponding

effect has no influence on network evolution. If the β_k is positive, the network tends to move towards the direction guided by the corresponding effect, and if β_k is negative, the network would "show resistance" to change into the direction indicating by the effect. [22] suggested the choice of effects included in the model can be guided by the observed data although some basic effects are almost always included in models, such as the density, reciprocity and transitivity.

The density of an actor is the number of friends she/he nominates. It reflects the tendency for an actor to establish a relationship with others in the network. Reciprocity measures an actor's tendency toward reciprocation of choices. Since friendship tends to follow the reciprocation norm, there is almost always a significant positive evidence for reciprocity in friendship network [24]. The actor-based model also allows complicated dependencies between ties such as the network's tendency toward transitivity. A "my friend's friend is my friend" situation is a common phenomenon, forming a triadic closure structure where two paths tend to become closed. The transitive triplets is a common measure of a network's transitivity by counting the number of the pattern (i, h, j), in which three actors are tied as $i \rightarrow j, j \rightarrow h, i \rightarrow h$. [22].

B. Simulation Design

Our simulation is based on networks generated from the actor-based model. Several factors that potentially influence clustering accuracy are considered in our simulation. First, we simulate networks with 2, 3, 5, and 10 clusters, respectively. Second, for each cluster, we consider three different sizes: 10, 20 and 30 actors. Therefore, for a network with 2 clusters each with size 10, there are a total of 20 actors in the network. Third, within each cluster, longitudinal social networks are simulated from the stochastic actor-based model using the method discussed in [22]. Two effects, reciprocity and transitivity, are considered in the objective function. The coefficients of the two effects are set at [3,3], [1,5], and [5,1], respectively. Fourth, 6 levels of cluster separation are considered: 0, 1%, 2%, 5%, 10%, and 20%. The value 0 means there is no between cluster ties, indicating actors from two clusters have no relationship at all. 20% means that 20% of all potential ties in a network is added randomly to the network. Therefore, 20% represents a condition with a significant amount of noise. Fifth, we also allow the ties among actors between clusters to endure over time. Specifically, such ties have the probability of 0.2, 0.5, and 1, respectively, to exist from one wave to the next wave. For convenience, we call this transitive probability. Sixth, the waves of data in the longitudinal network are 2, 3, 5, and 10, respectively. Based on the 6 factors, for each combination (4 number of clusters $\times$ 3 different sizes $\times$ 3 sets of coefficients $\times$ 6 levels of separations $\times$ 3 different transitive probability $\times$ 4 different waves = 2592 combinations in total) of their levels, we generate 100 longitudinal social networks and use our clustering method to identify the clusters.

C. Results

To analyze the influence of the 6 factors on the clustering accuracy, we first conducted an ANOVA analysis with the clustering accuracy as the outcome. The result showed that the number of clusters, the size of the network, the level of cluster separation and the number of waves were significantly related to the clustering accuracy. On the other hand, the transitive probability and the effect coefficients were insignificant. In order to show the effect of each factor graphically, we varied one factor but fixed the other factors to be the same. Then, we sorted the average accuracy and plotted them according to the different levels of the factor of interest.

The following conclusions can be drawn from our simulation study. First, for the majority of conditions, the clustering accuracy was high. For example, in more than 85% of the 2592 conditions, the clustering accuracy was greater than 90%, indicating 90% of the actors were clustered into the desired cluster. Only in about 5% of the conditions, the clustering accuracy was lower than 80%. Second, the clustering accuracy did not seem to vary much according to the transitive probability and the effect coefficients. Third, the number of clusters was negatively related to the accuracy. With more clusters, it became more difficult to cluster the actors. Fourth, our clustering method was more accurate with smaller networks than bigger networks. Fifth, the clustering accuracy was high when the cluster separation was high. For example, when the class separation is 0, indicating complete separation, the accuracy was almost 1 all the time. Sixth, with the increase of the number of waves, the clustering accuracy became better.

IV. Real Data Analysis

To illustrate the application of our clustering method, we use a network of college students, which was initially studied by [25]. The data were collected in 1994 and 1995, started with 56 students at a university in the Netherlands. The students were asked to answer a questionnaire 7 times throughout an academic year. Some students dropped out during the academic year, and were removed from the data set. Students replying to the questionnaire less than 4 times were also removed, which led to a network with 32 students. In addition, one student reported no friendship relation with any other students in the network and was also removed in the current analysis. Therefore, the total number of students in the real data analysis is 31. The estimated strength of friendship among any pair of students was put together into a matrix, which is displayed in Figure 1. In the figure, the darker color represents stronger friendship. The estimated strength ranged from 0.07 to 0.95. This matrix is used to group students into different clusters.

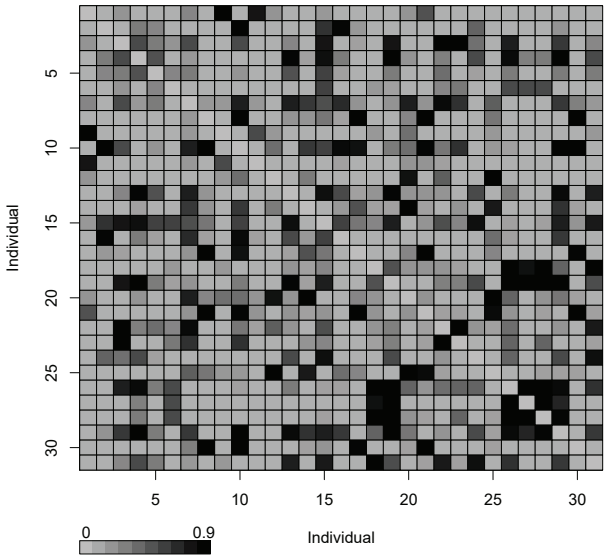


Figure 1. Strength matrix estimated for the 31 students from the 7 waves of data

Firstly, we decided on the number of clusters based on the Calinski-Harabasz Criterion [26], which performed best among a set of criteria in the simulation carried out by [27]. According to the Calinski-Harabasz Criterion, retaining two clusters provides the largest Calinski-Harabasz value. Therefore, we clustered the network into two clusters, one was 10-centered, which means the 10th actor was the focus of this group, and the other was 19-centered. The two clusters are also shown in Figure 2. The 10-centered cluster contained 17 actors and the 19-centered cluster consisted of 14 actors.

To illustrate the potential structural difference between the two clusters, we fit the actor-based model to each cluster and the overall data. In the model, three effects were considered: density, reciprocity, and popularity. The estimated coefficients for the three effects are shown in Table II. For the first cluster, density effect and reciprocity effect were significant, but the popularity effect was insignificant. For the second cluster group, all three effects were significant. Furthermore, the direction of the density effect was opposite for the two clusters. When fitting the model to all data, reciprocity effect and popularity effect were significant while the density effect was not. The results showed the structural difference between the two clusters and also justified the existence of the two clusters. The two clusters are shown in Figure 2(a).

In the literature, other strategies have been used to cluster longitudinal networks. For example, for one method, one can cluster the networks based on the static, last wave of data. For another, one can first aggregate the longitudinal networks together and then cluster the aggregated network [28]. For comparison, we also applied the two methods to this set of real data.

Table II
RESULTS FROM THE ACTOR-BASED MODEL FOR EACH CLUSTER AND OVERALL NETWORK

		Estimate	Standard Error	Sig. Level
Clustering based on tie strength				
Cluster 1	Density	-1.61	0.363	***
	Reciprocity	2.21	0.226	***
	Popularity	-0.22	0.191	
Cluster 2	Density	1.26	0.545	***
	Reciprocity	2.27	0.312	***
	Popularity	-0.78	0.258	***
Clustering based on the last wave of data				
Cluster 1	Density	2.06	1.185	
	Reciprocity	3.32	0.832	***
	Popularity	-1.33	0.649	*
Cluster 2	Density	-0.42	0.340	
	Reciprocity	2.47	0.205	***
	Popularity	-0.66	0.198	***
All data				
	Density	-0.33	0.200	
	Reciprocity	2.23	0.122	***
	Popularity	-0.38	0.088	***

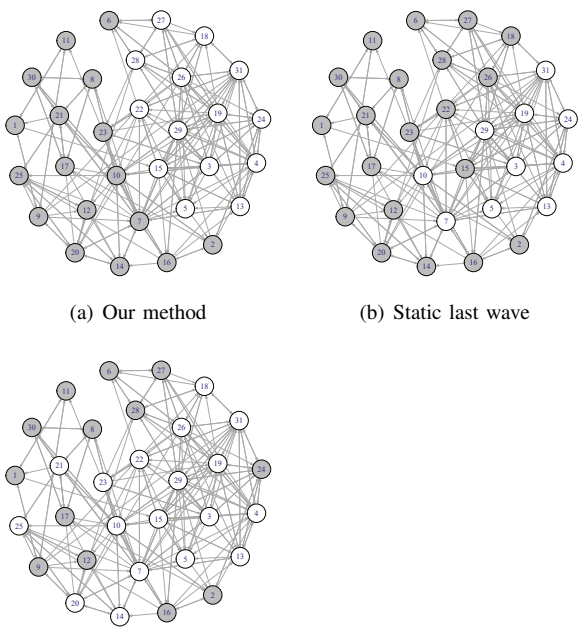
Note. * significant at alpha level 0.05, *** significant at alpha level 0.001.

The two clusters based on the last wave of the static network are displayed in Figure 2(b) and based on the aggregated network are displayed in Figure 2(c). Visually, we can see that there exists a difference in the members of each clustering strategy. Furthermore, the Cohen's κ [29] is 0.47 between our method and the results from the last wave of the network, and 0.36 between our method and the aggregated network, respectively. Therefore, the agreement in clustering is only fair to moderate [30].

We also fitted the actor-based model to each cluster identified using either the last wave of network or the aggregated network. The model did not converge for the aggregated network and therefore no results were reported. The results based on the last wave of data are shown in Table II. For both clusters, the density effect was not significant but the reciprocity and popularity effects were significant, similar to the use of all students' data. Therefore, clustering only based on the last wave of data showed a less clear difference between the two clusters comparing to the use of our method based on the tie strength.

V. CONCLUSION AND DISCUSSION

In this paper, we proposed a new method for longitudinal social network clustering. Previous researches on the clustering of longitudinal networks mainly focused on investigating how clusters form, evolve and die by analyzing the time snapshots of a dynamic network [2], [31], [32]. [33] kept track of the clusters by analyzing the nodes and the cluster they belonged to. [34] reformulated the known clustering techniques for a static network to the framework for a longitudinal network by incorporating temporal smoothness.



(a) Our method (b) Static last wave

(c) Aggregated network

Figure 2. The plots of the clusters identified by different methods.

Our method focused on estimating the strength of the ties according to the evolution throughout time. Once the strength is estimated, the existing clustering methods can be applied such as the spectral clustering method used in the current study.

A simulation study showed that our method can identify the correct clusters in most conditions evaluated. It also showed that the clustering accuracy was influenced by the number of clusters, the size of the network, the level of cluster separation and the number of waves. Furthermore, we identified two clusters in the data collected by [25], which are widely used for longitudinal network analysis. The analysis of each cluster through the actor-based model indicated the structural difference in the two clusters.

Our method has many advantages. First, the strength among ties can be estimated easily. The Bayesian estimation has a simple form as the ratio of frequencies. Second, the Bayesian method can effectively deal with extreme cases. Because of the use the prior, even uninformative, the extreme cases can be adjusted to be more realistic. Third, once a strength matrix is formed, many existing clustering algorithms can be applied. In the current study, the k-medoids clustering method was used. However, other methods can be applied as well.

Our method can be expanded in many ways in the future. First, we applied our method to the actor-based network data in the current study. However, the same idea can be applied to longitudinal networks generated using other mechanisms. Second, the basic spectral clustering algorithm was used after we obtained the weighted adjacency matrix. Future studies can investigate better clustering algorithms that may lead to better performance when applying our method. Third, social network imputation has been studied in the literature [35], [36] and our algorithm can also be applied since we can estimate the distribution of the Markov Chain. Fourth, we can consider effects such as reciprocity when evaluating the strength of the ties and this might provide more accurate clustering.

Like any other method, our proposed method is not without caveats. First, in estimating the tie strength of a pair of actors, we did not take into account the potential relationship with or dependence on other actors. A potential future study is to investigate how to estimate a partial tie strength by removing the effect of other actors. Second, the current study has focused on the use of the actor-based model to illustrate the use of the proposed method. Many other kinds of social networks such as the ones based on the actor-level model can be investigated in the future. Third, we only evaluated one way to measure the tie strength. Other methods can be explored in the future. Even with these caveats, we have shown that our method can perform well both in simulation and real data situations. We hope that our method can contribute to the growing body of approaches for longitudinal network analysis.

REFERENCES

[1] J. Scott, *Social network analysis*. Sage, 2012.

[2] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010.

[3] Y. Kim, S.-W. Son, and H. Jeong, "Finding communities in directed networks," *Physical Review E*, vol. 81, no. 1, p. 016103, 2010.

[4] M. Meilă and W. Pentney, "Clustering by weighted cuts in directed graphs," in *SIAM Conference on Data Mining (SDM). Retrieved February*, vol. 1. SIAM, 2007, p. 2008.

[5] M. C. Nascimento and A. C. De Carvalho, "Spectral methods for graph clustering–a survey," *European Journal of Operational Research*, vol. 211, no. 2, pp. 221–231, 2011.

[6] V. Satuluri and S. Parthasarathy, "Symmetrizations for clustering directed graphs," in *Proceedings of the 14th International Conference on Extending Database Technology*. ACM, 2011, pp. 343–354.

[7] D. Duan, Y. Li, Y. Jin, and Z. Lu, "Community mining on dynamic weighted directed graphs," in *Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management*. ACM, 2009, pp. 11–18.

[8] F. D. Malliaros and M. Vazirgiannis, "Clustering and community detection in directed networks: A survey," *Physics Reports*, vol. 533, no. 4, pp. 95–142, 2013.

[9] T. A. Snijders, M. Spreen, and R. Zwaagstra, "The use of multilevel modeling for analysing personal networks: Networks of cocaine users in an urban area," *Journal of quantitative anthropology*, vol. 5, no. 2, pp. 85–105, 1995.

[10] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," *Journal of Machine Learning Research*, vol. 9, no. Sep, pp. 1981–2014, 2008.

[11] L. C. Freeman, "The sociological concept of" group": An empirical test of two models," *American journal of sociology*, pp. 152–166, 1992.

[12] P. D. Hoff, "Bilinear mixed-effects models for dyadic data," *Journal of the american Statistical association*, vol. 100, no. 469, pp. 286–295, 2005.

[13] P. F. Lazarsfeld and N. W. Henry, *Latent structure analysis*. Houghton Mifflin Co., 1968.

[14] T. A. Snijders, "Statistical models for social networks," *Annual Review of Sociology*, vol. 37, pp. 131–153, 2011.

[15] S. Uddin, A. Khan, L. Hossain, M. Piraveenan, and S. Carlsson, "A topological framework to explore longitudinal social networks," *Computational and Mathematical Organization Theory*, vol. 21, no. 1, pp. 48–68, 2015.

[16] S. Uddin, A. Khan, and M. Piraveenan, "A set of measures to quantify the dynamicity of longitudinal social networks," *Complexity*, vol. 21, no. 6, pp. 309–320, 2016.

[17] W. E. Donath and A. J. Hoffman, "Lower bounds for the partitioning of graphs," *IBM Journal of Research and Development*, vol. 17, no. 5, pp. 420–425, 1973.

[18] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak mathematical journal*, vol. 23, no. 2, pp. 298–305, 1973.

[19] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.

[20] S. M. Ross *et al.*, *Stochastic processes*. John Wiley & Sons New York, 1996, vol. 2.

[21] H. Jeffreys, "An invariant form for the prior probability in estimation problems," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 186, no. 1007. The Royal Society, 1946, pp. 453–461.

[22] T. A. Snijders, G. G. Van de Bunt, and C. E. Steglich, "Introduction to stochastic actor-based models for network dynamics," *Social networks*, vol. 32, no. 1, pp. 44–60, 2010.

[23] T. A. Snijders, "Stochastic actor-oriented models for network change," *Journal of mathematical sociology*, vol. 21, no. 1-2, pp. 149–172, 1996.

[24] R. Veenstra, J. K. Dijkstra, C. Steglich, and M. H. Van Zalk, "Network–behavior dynamics," *Journal of Research on Adolescence*, vol. 23, no. 3, pp. 399–412, 2013.

[25] G. G. Van de Bunt, M. A. Van Duijn, and T. A. Snijders, "Friendship networks through time: An actor-oriented dynamic statistical network model," *Computational & Mathematical Organization Theory*, vol. 5, no. 2, pp. 167–192, 1999.

[26] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[27] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[28] S. Uddin, K. S. K. Chung, and M. Piraveenan, "Capturing actor-level dynamics of longitudinal networks," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, 2012, pp. 1006–1011.

[29] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[30] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[31] S. Asur and S. Parthasarathy, "An event-based framework for characterizing the evolution of interaction graphs," 2007.

[32] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, "Tracking evolving communities in large linked networks," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5249–5253, 2004.

[33] D. J. Fenn, M. A. Porter, M. McDonald, S. Williams, N. F. Johnson, and N. S. Jones, "Dynamic communities in multichannel data: An application to the foreign exchange market during the 2007–2008 credit crisis," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 19, no. 3, p. 033119, 2009.

[34] Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng, "Evolutionary spectral clustering by incorporating temporal smoothness," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 153–162.

[35] M. Huisman, "Imputation of missing network data: some simple procedures," *Journal of Social Structure*, vol. 10, no. 1, pp. 1–29, 2009.

[36] A. Žnidaršič, A. Ferligoj, and P. Doreian, "Non-response in social networks: The impact of different non-response treatments on the stability of blockmodels," *Social Networks*, vol. 34, no. 4, pp. 438–450, 2012.