A Longitudinal Social Network Clustering Method Based on Tie Strength

Zhiyong Zhang, University of Notre Dame
Ye Mao, Indiana University
Yijie Huang, Columbia University
Nan Sun, Sichuan University

2018 IEEE International Conference on Big Data
Dec 11, 2018 | Seattle, WA

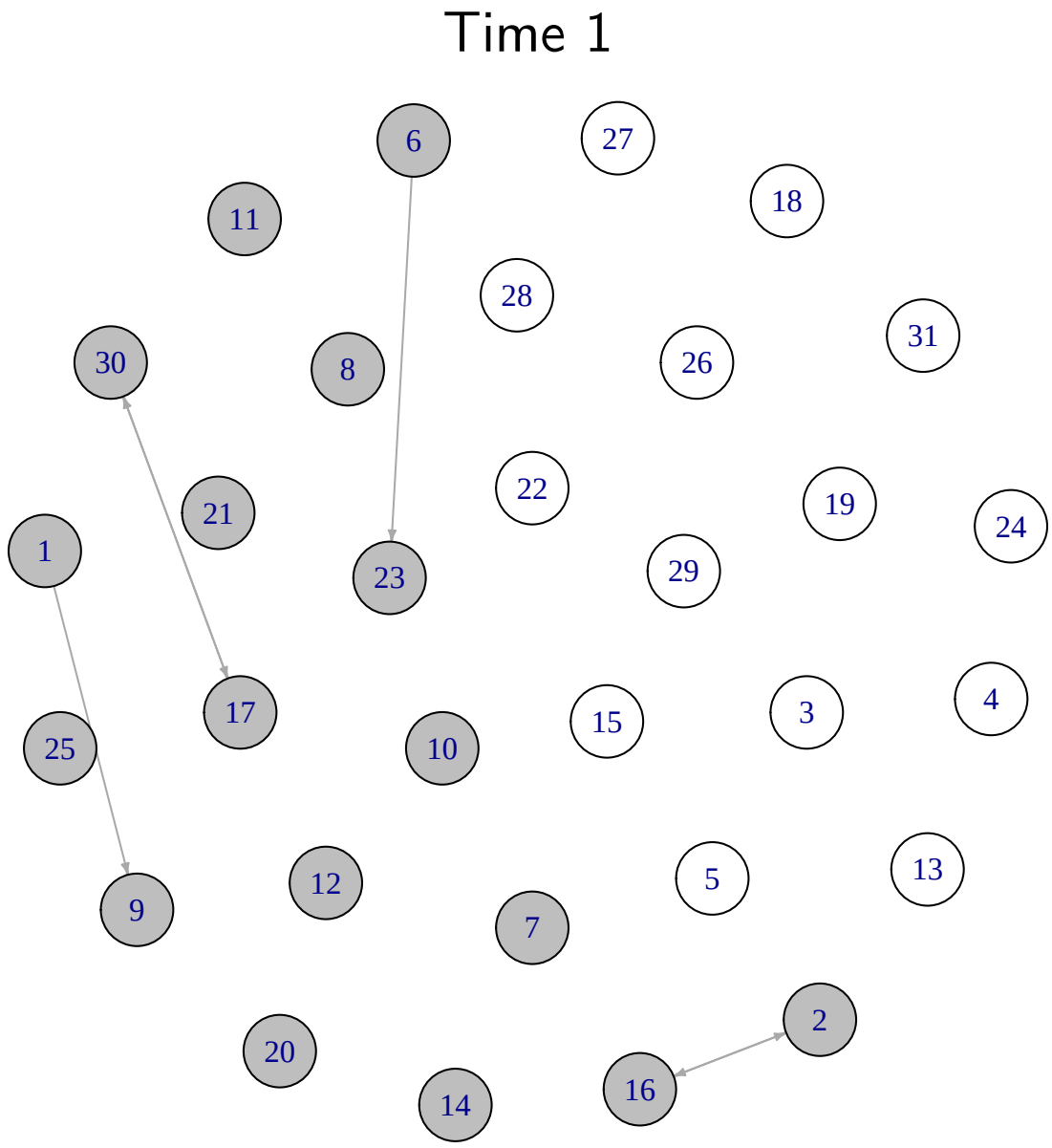
Outline

- Data
- Existing clustering methods
- Our proposed method
- Simulation study
- An application

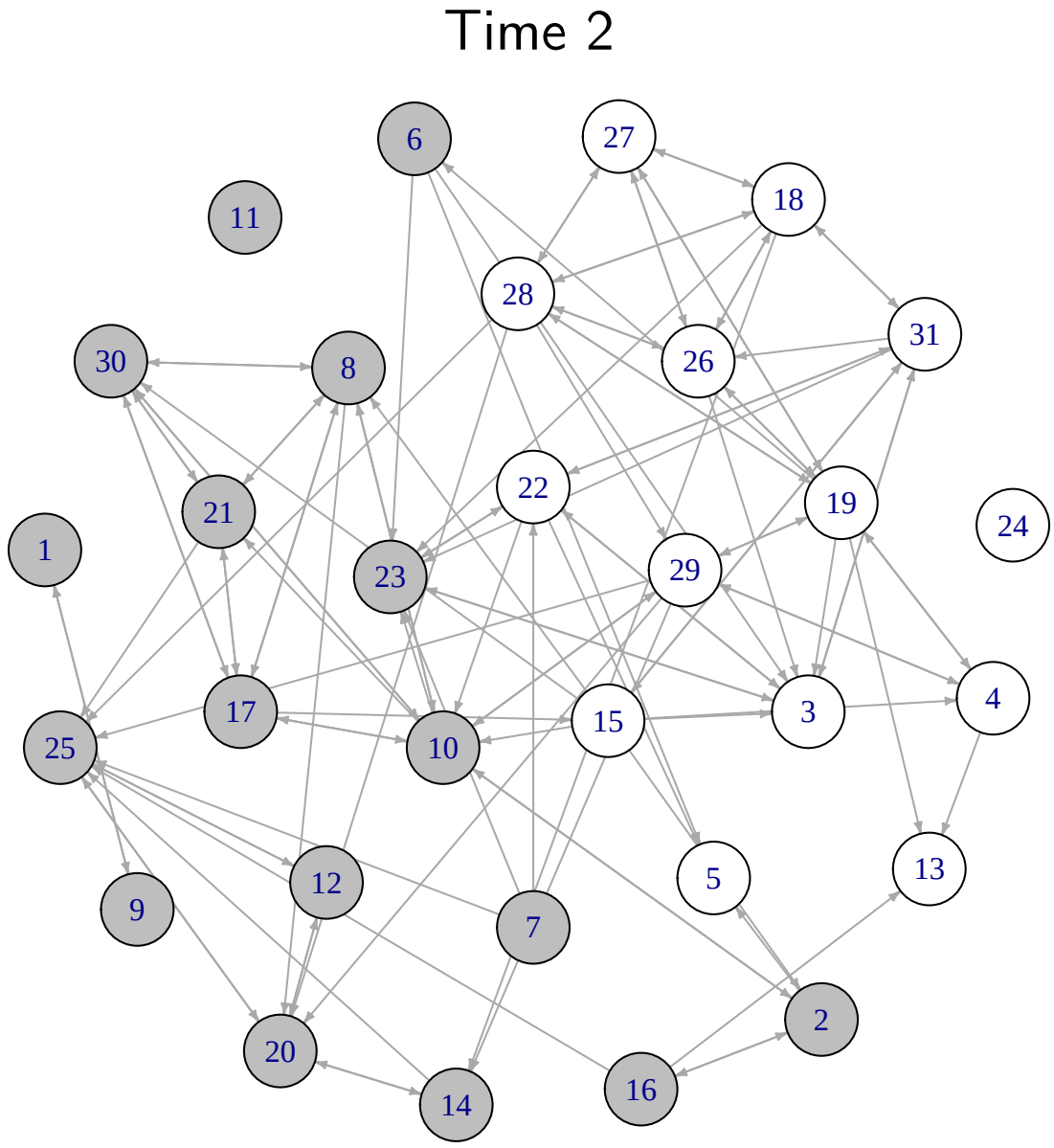
An example data set on friendship

- The data were collected by Gerhard van de Bunt among a group of university freshmen.
- A total of 7 times of data were collected. At the first measurement, most students did not know each other
- The first four time points are three weeks apart, whereas the last three time points are six weeks apart.
- The original group consisted of 49 students, but due to 'university drop-outs' and after deleting those who did not fill in the questionnaire four or more times, a group of 32 students were left.
- One student reported no friendship relation with any other students in the network and was also removed.
- The sample in this study has 31 students.

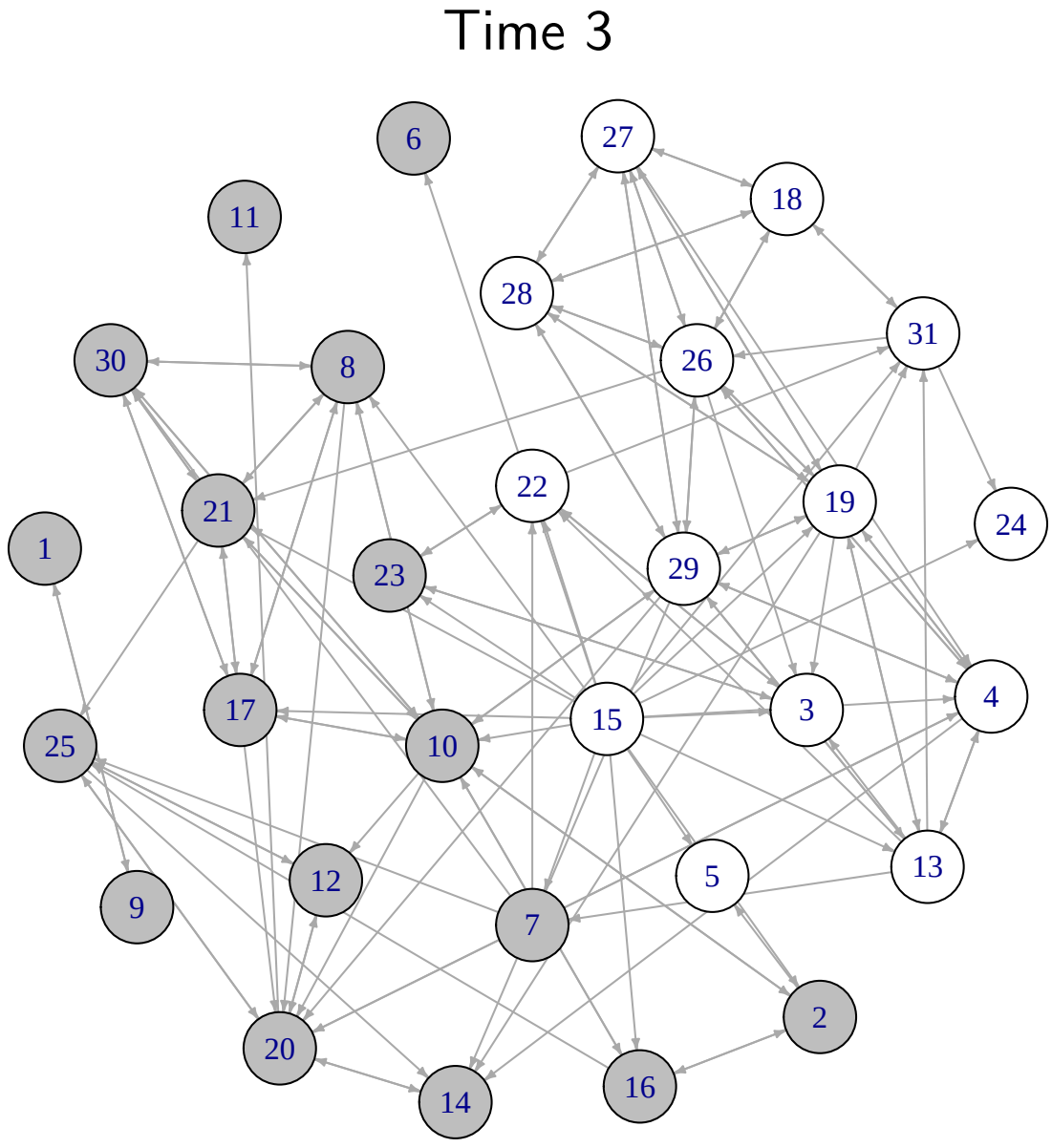
Longitudinal network plot



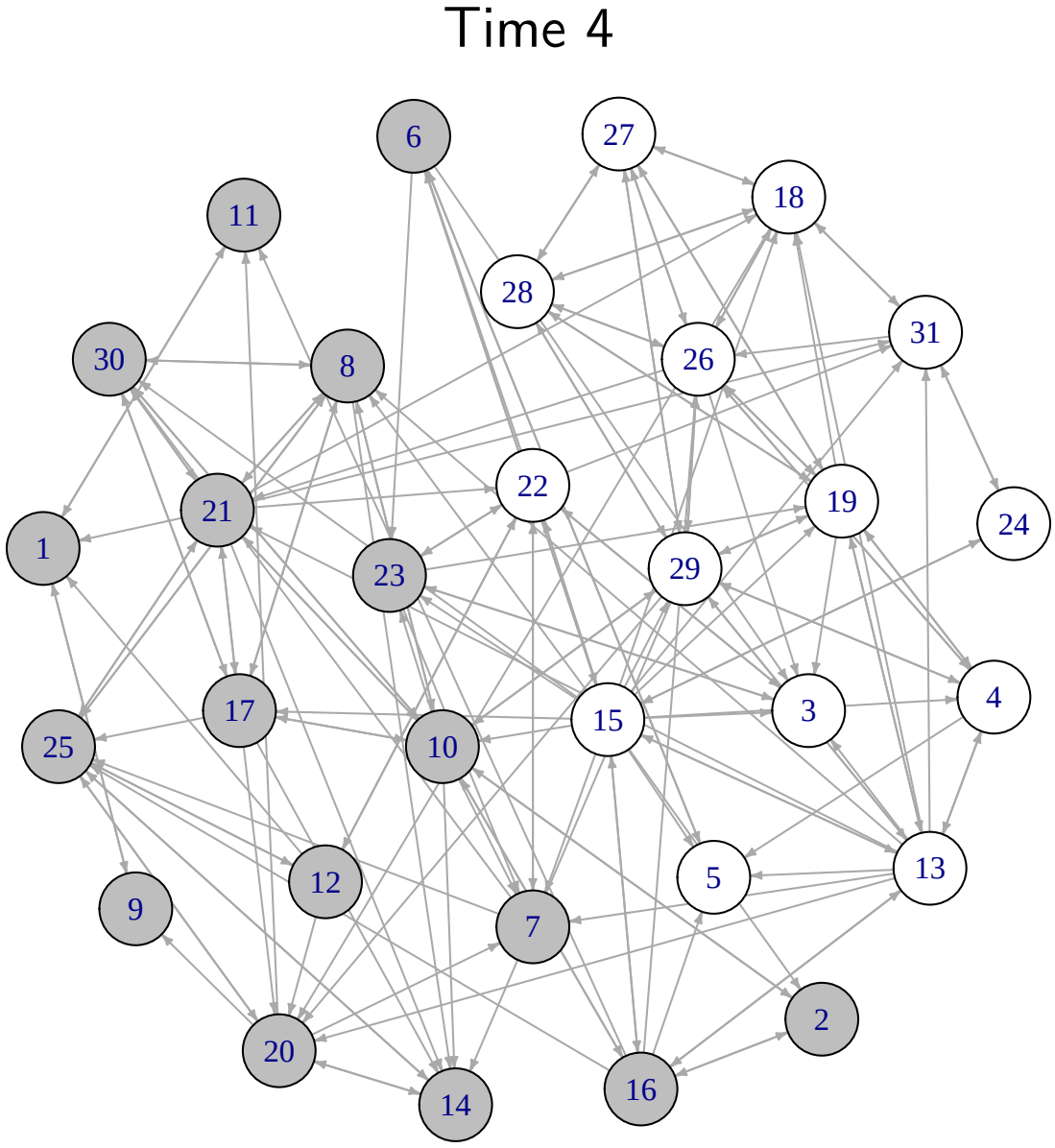
Longitudinal network plot



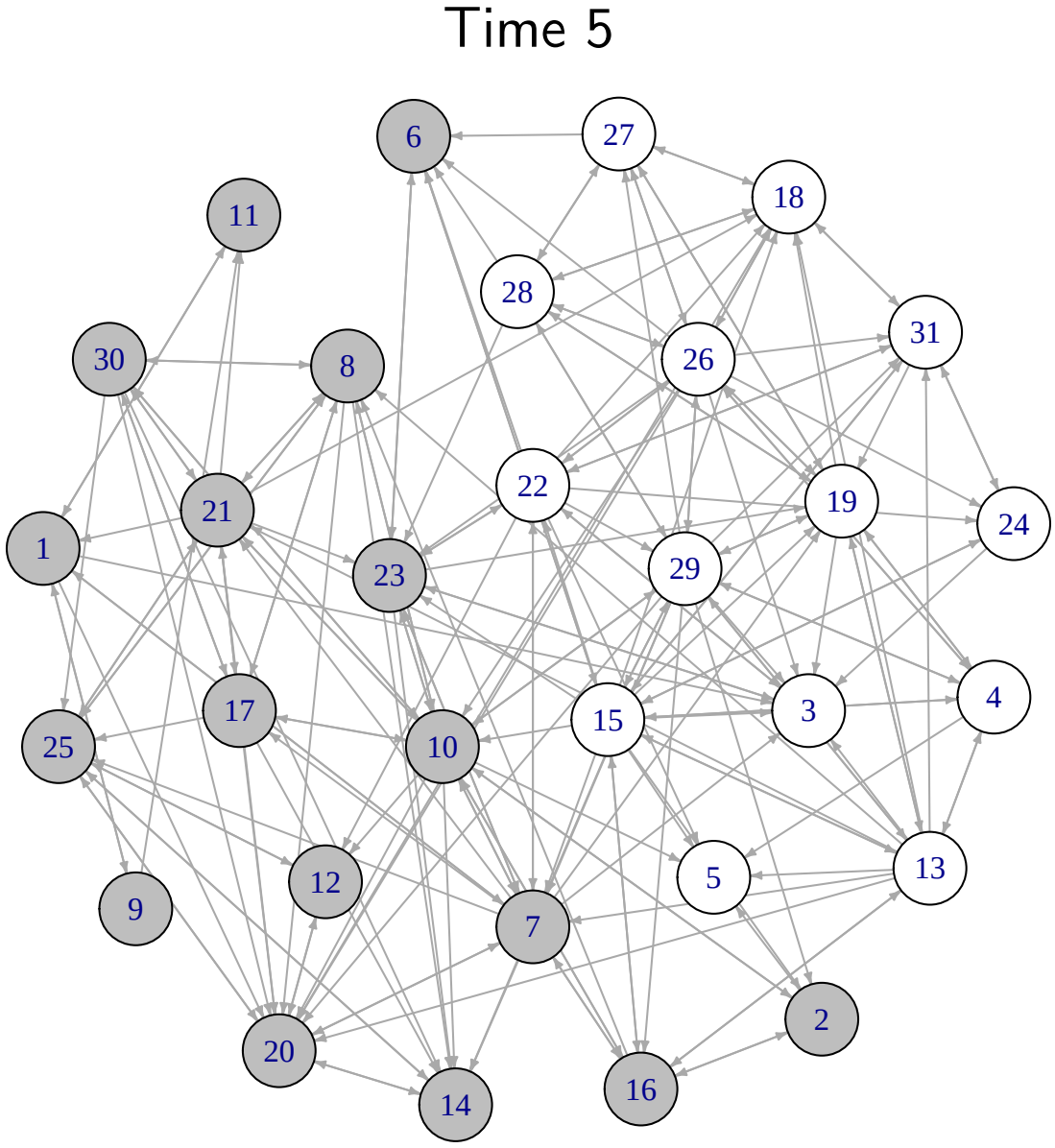
Longitudinal network plot



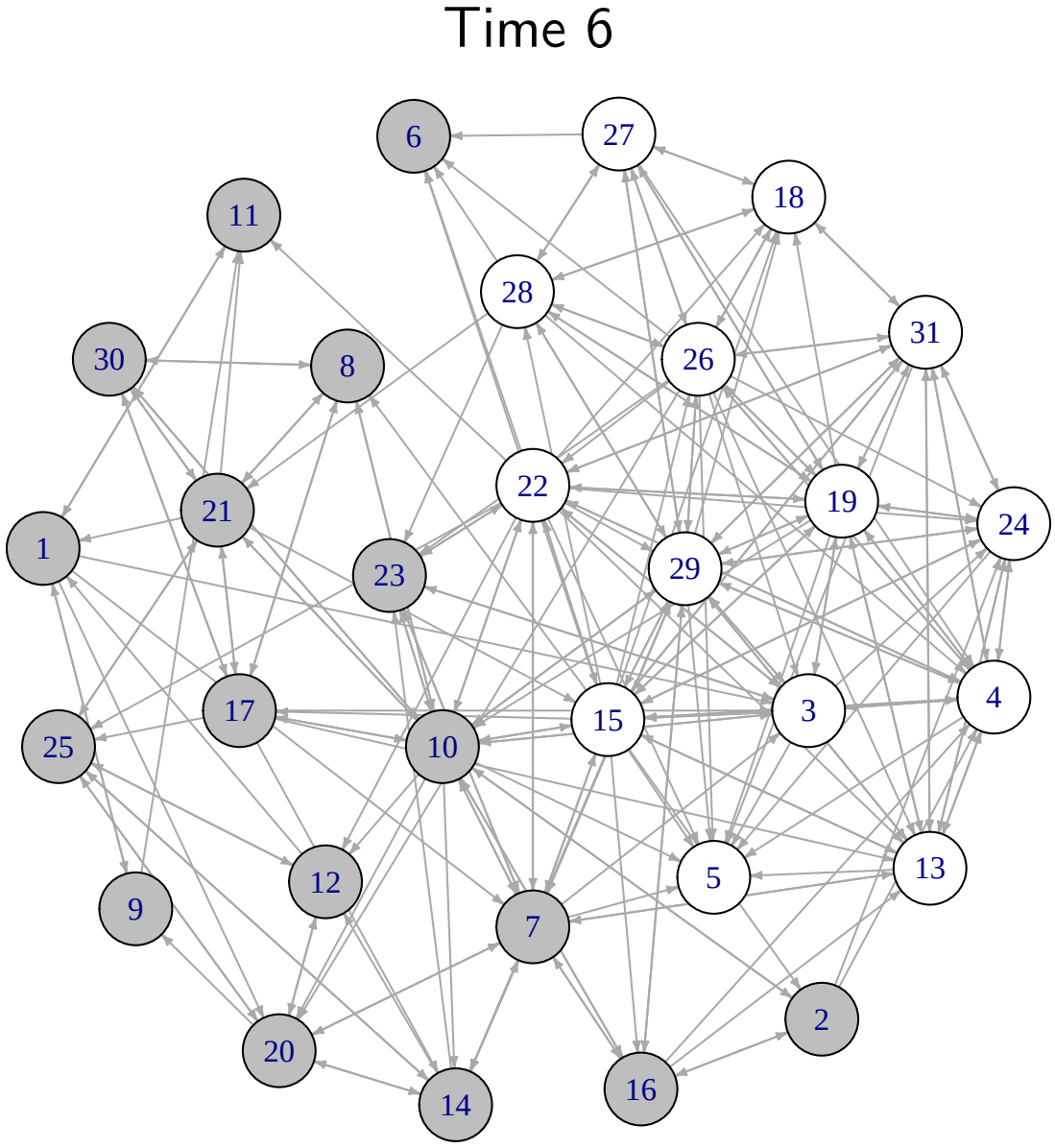
Longitudinal network plot



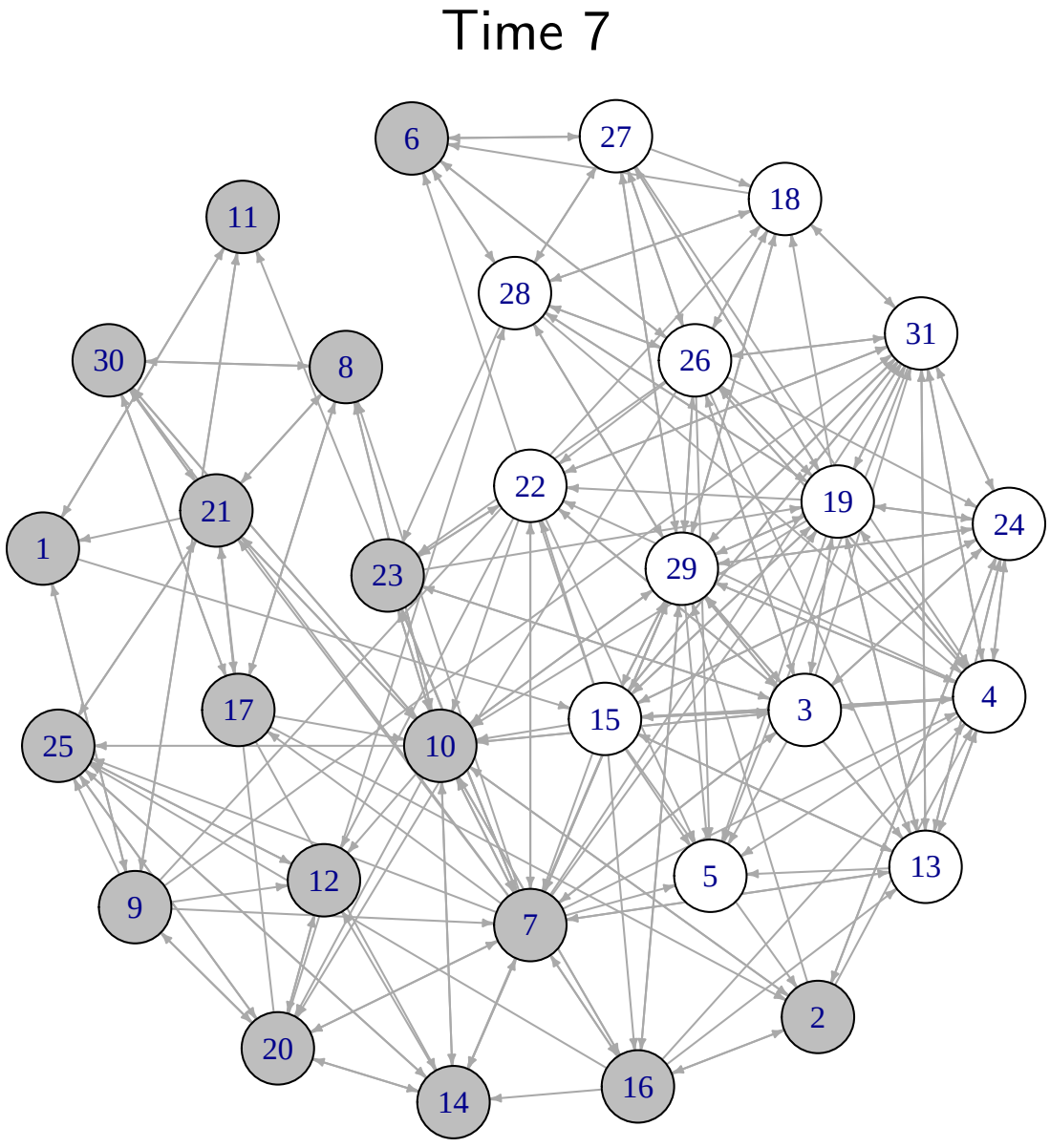
Longitudinal network plot



Longitudinal network plot



Longitudinal network plot



Longitudinal network clustering: existing methods

- Identifying the clusters based on the static snapshots of the longitudinal network, and detecting the change points throughout the time according to the different partitions over time (Malliaros and Vazirgiannis, 2013).
- Clustering the networks based on the static, last wave of data.
- Aggregating the longitudinal networks together and then clustering the aggregated network (e.g., Uddin et al., 2012).
- The existing methods have focused on networks in which ties are regarded as brief events such as email and cell-phone communications among actors (Duan et al., 2009; Sun et al., 2007).
- In many social networks on friendship, trust, and cooperation, a tie or relationship often builds gradually and endures over time as opposed to be created and terminated spontaneously.

Our proposed method

- 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.
- The general idea is to evaluate the strength of the ties according to how a network evolves throughout time.
- We assume that in a small time interval there is an overwhelming probability that the state will remain unchanged; however, if it changes, the change may be radical.
- One way to describe the process is to use the Kolmogorov forward equation.

Define a tie strength measure

- Let $P_{lk}(t)$ be the probability for the Markov process transferring from state l to state k, $l, k = 0, 1$.
- Solving the two-state Kolmogorov forward equation

$$P'_{00}(t) = \lambda_1 P_{01}(t) - \lambda_0 P_{00}(t) = -(\lambda_0 + \lambda_1) P_{00}(t) + \lambda_1$$

leads to the transition probability

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

  - λ_0 is the rate of expential time spending on state 0 before moving to 1.
  - λ_1 is the rate of expential time spending on state 1 before moving to 0.
- We propose the following tie strength measure

$$\theta_1 = \lim_{t \to \infty} P_{11}(\mathrm{t}) = \frac{\lambda_0}{\lambda_1 + \lambda_0}.$$

A Bayesian estimator I

- From the transition probablity, we have

$$\theta_1 = \frac{\lambda_0}{\lambda_1 + \lambda_0} = \frac{P_{01}}{1 - P_{11} + P_{01}}.$$

- Define

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

  - $X_{ij}(t_l)$ is a pair of nodes at time t_l from the longitudinal network
  - M is the total number of times
  - $h, l = 0, 1$

- An estimation of the tie strength is

$$\hat{\theta}_1 = \frac{\widehat{\lambda_0}}{\lambda_0 + \lambda_1} = \frac{\frac{\frac{1}{2}+N_{01}}{1+N_{00}+N_{01}}}{1 - \frac{\frac{1}{2}+N_{01}}{1+N_{00}+N_{01}} + \frac{\frac{1}{2}+N_{11}}{1+N_{10}+N_{11}}}.$$

Spectral clustering

- The strength of the relationship between any two actors i and j in a network can be estimated using our Bayesian method to form an undirected strength matrix $\mathbf{W}$.
- Any existing methods for valued network clustering can be used to detect clusters among the actors.
- We use the spectral clustering for this study.
  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.

Simulation study

- To evaluate the performance of our method, we conduct a simulation study based on the actor-based model (e.g., Snijders et al., 2010; Snijders, 1996).
- The actor-based model is defined by its objective function

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

  - The function $s_{ik}(x)$ represents a component, or an effect, of the network, indicating a specific tendency of change in the network.
  - The parameter β_k is the weight of each effect.
  - Some basic effects include density, reciprocity and transitivity.
    - The density of an actor is the number of friends she/he nominates.
    - Reciprocity measures an actor's tendency toward reciprocation of choices.
    - Transitivity refers to the extent to which the relation that relates two nodes in a network that are connected by an edge is transitive – "my friend's friend is my friend."

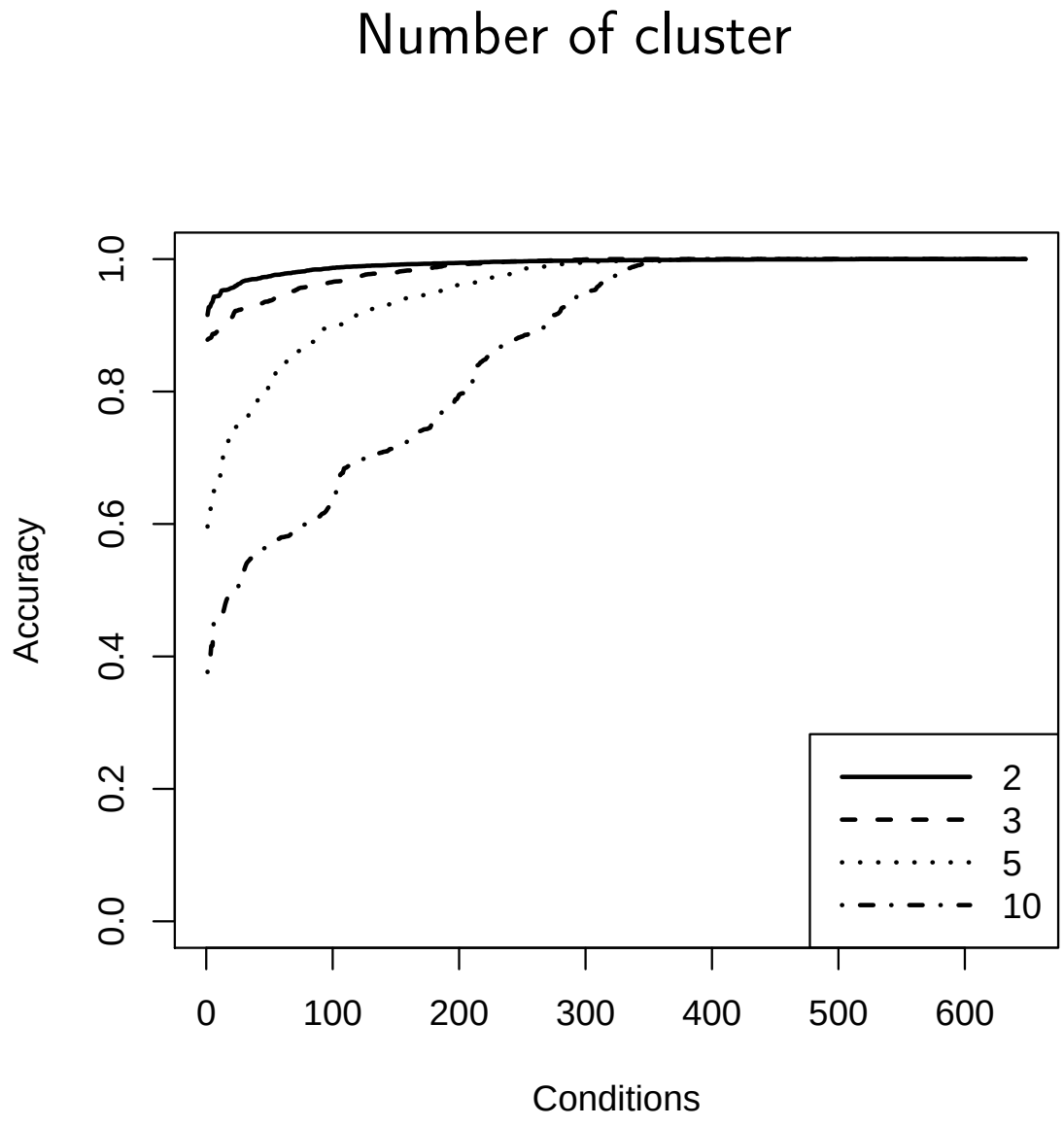
Simulation design

- Number of clusters: 2, 3, 5, and 10
- Cluster size: 10, 20 and 30 actors each cluster.
- Waves of the longitudinal network: 2, 3, 5, and 10
- Cluster separation: 0, 1%, 2%, 5%, 10%, and 20%. 0 means there is no between cluster ties, and 20% means that 20% of all potential ties in a network is added randomly.
- Coefficients of effects: [3,3], [1,5], and [5,1].
- Tie transitive probability: 0.2, 0.5, and 1. For convenience, we call this transitive probability.
- A total of 2592 conditions with 100 longitudinal social networks each.

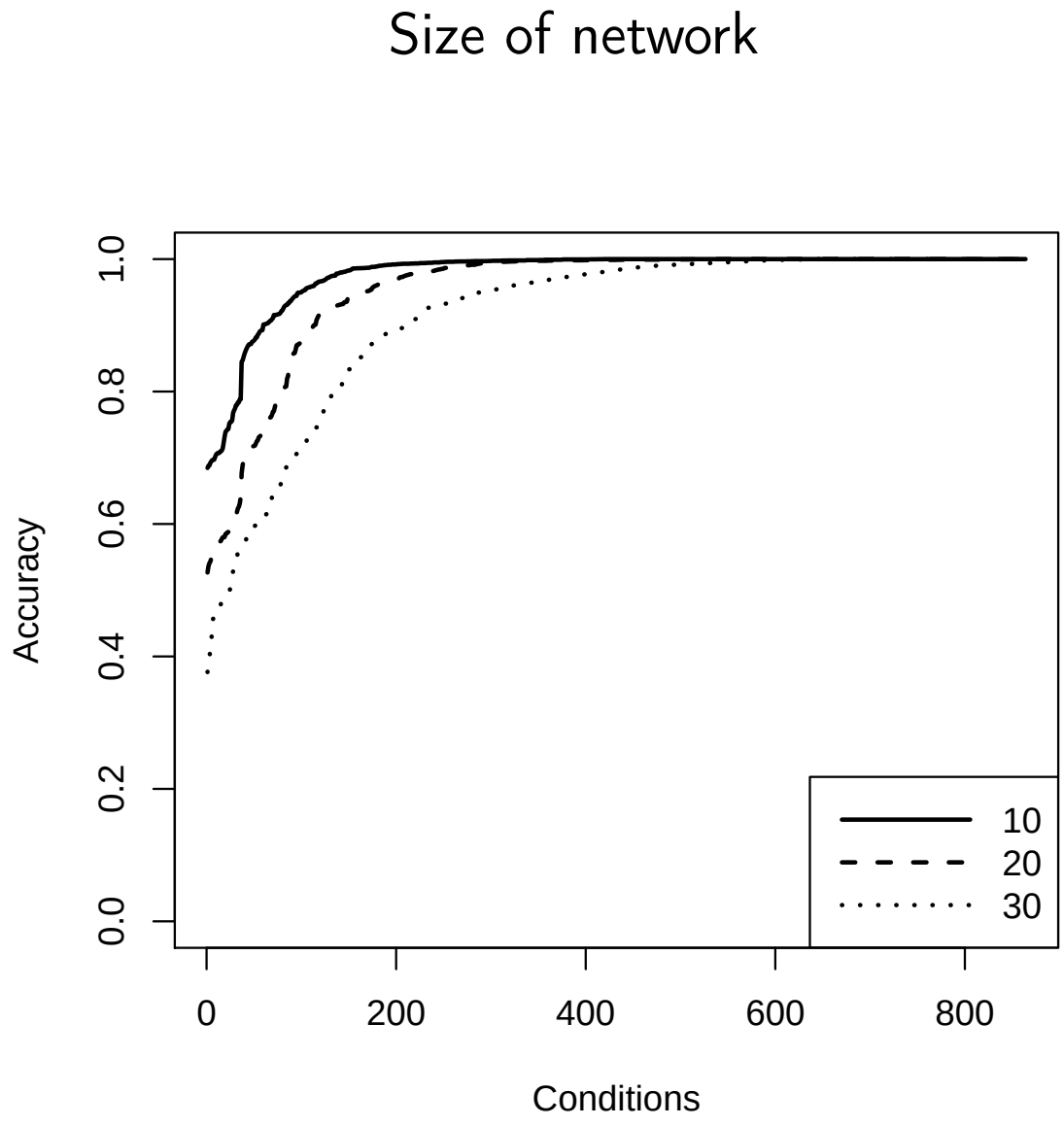
Simulation results

- For the majority of conditions, the clustering accuracy was high.
  - 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%.
- The clustering accuracy did not seem to vary much according to the transitive probability and the effect coefficients.
- The number of clusters was negatively related to the accuracy.
- The clustering method was more accurate with smaller networks than bigger networks.
- The clustering accuracy was high when the cluster separation was high.
- With the increase of the number of waves, the clustering accuracy became better.

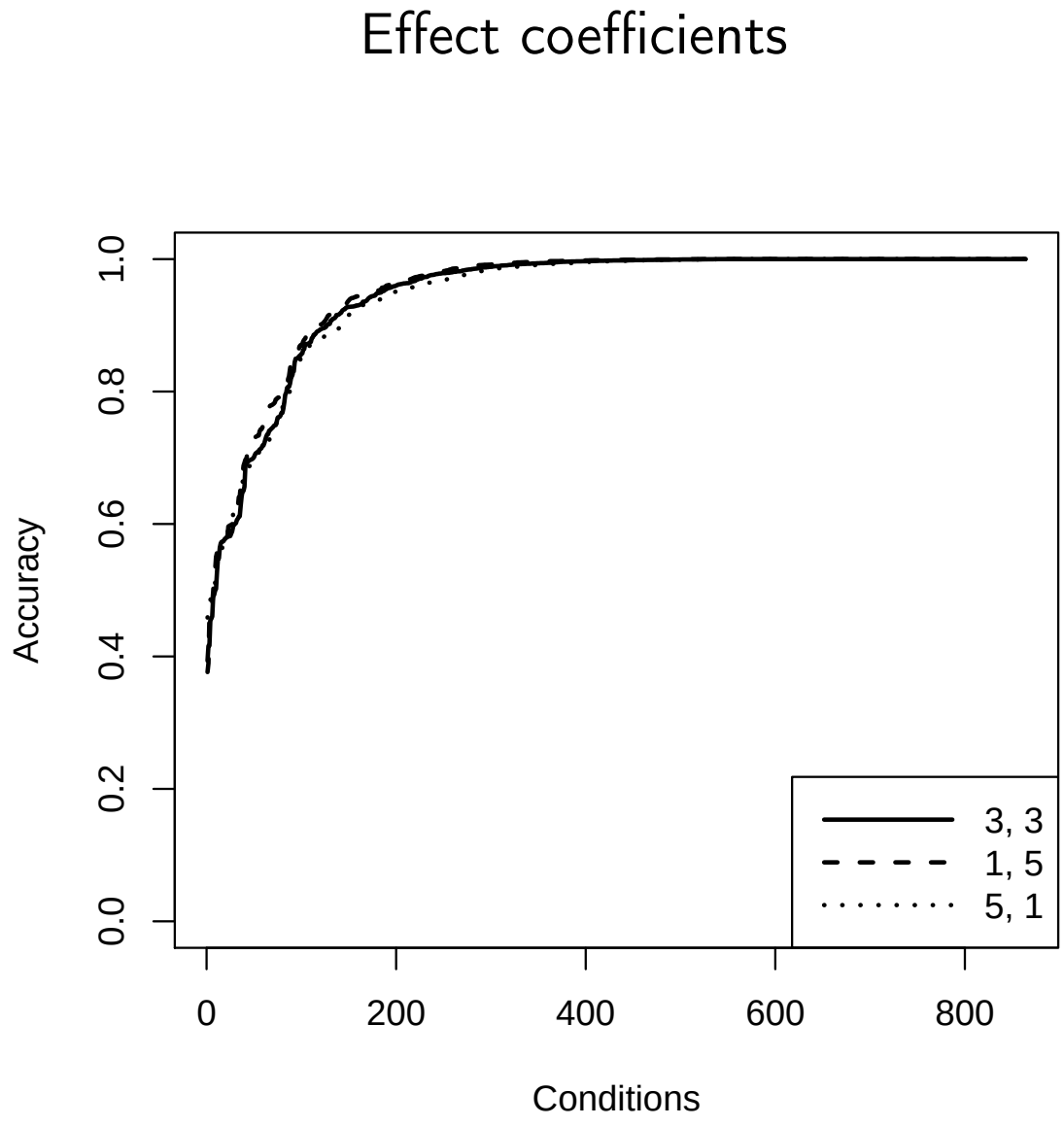
Influence of each factor



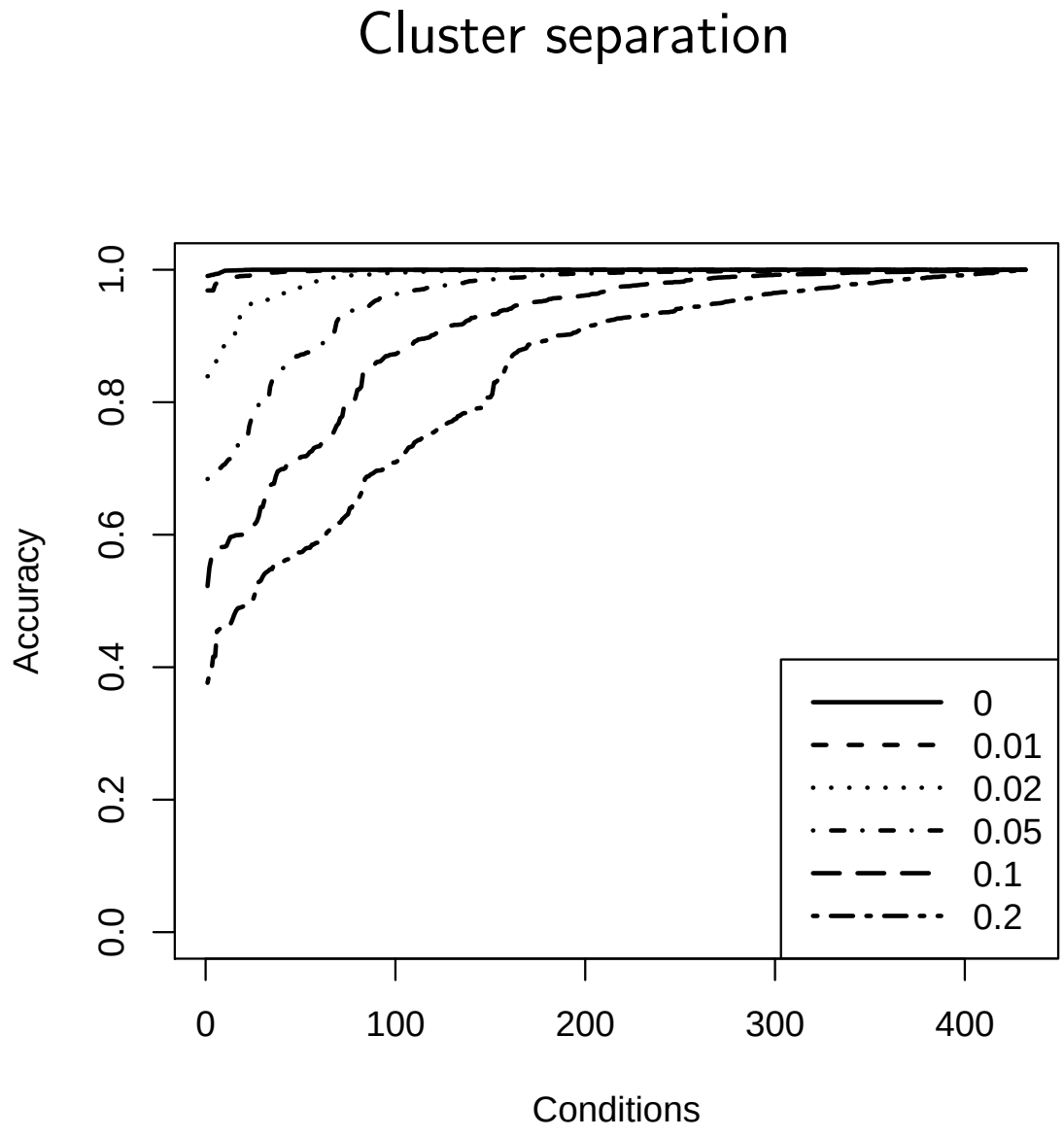
Influence of each factor



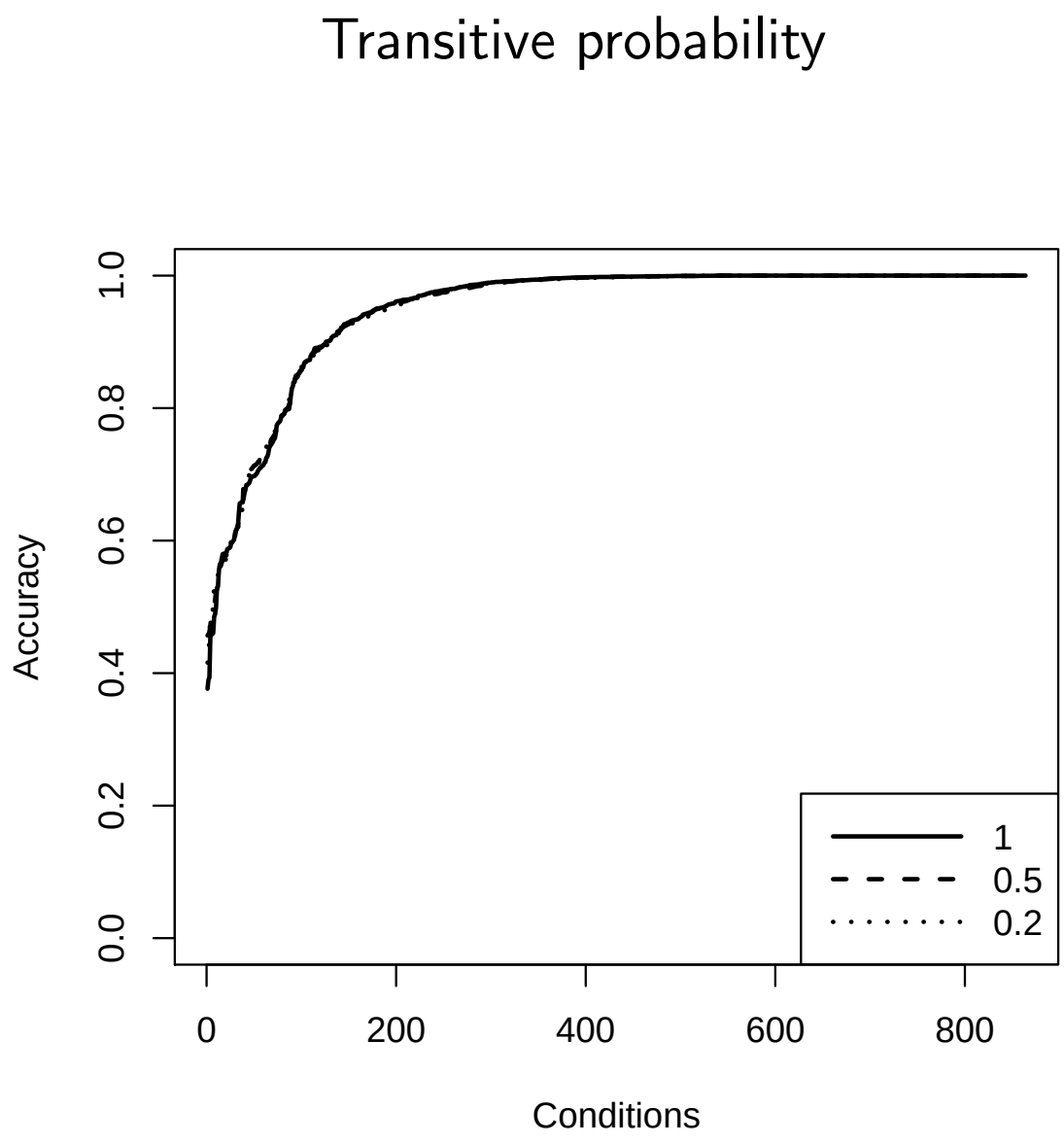
Influence of each factor



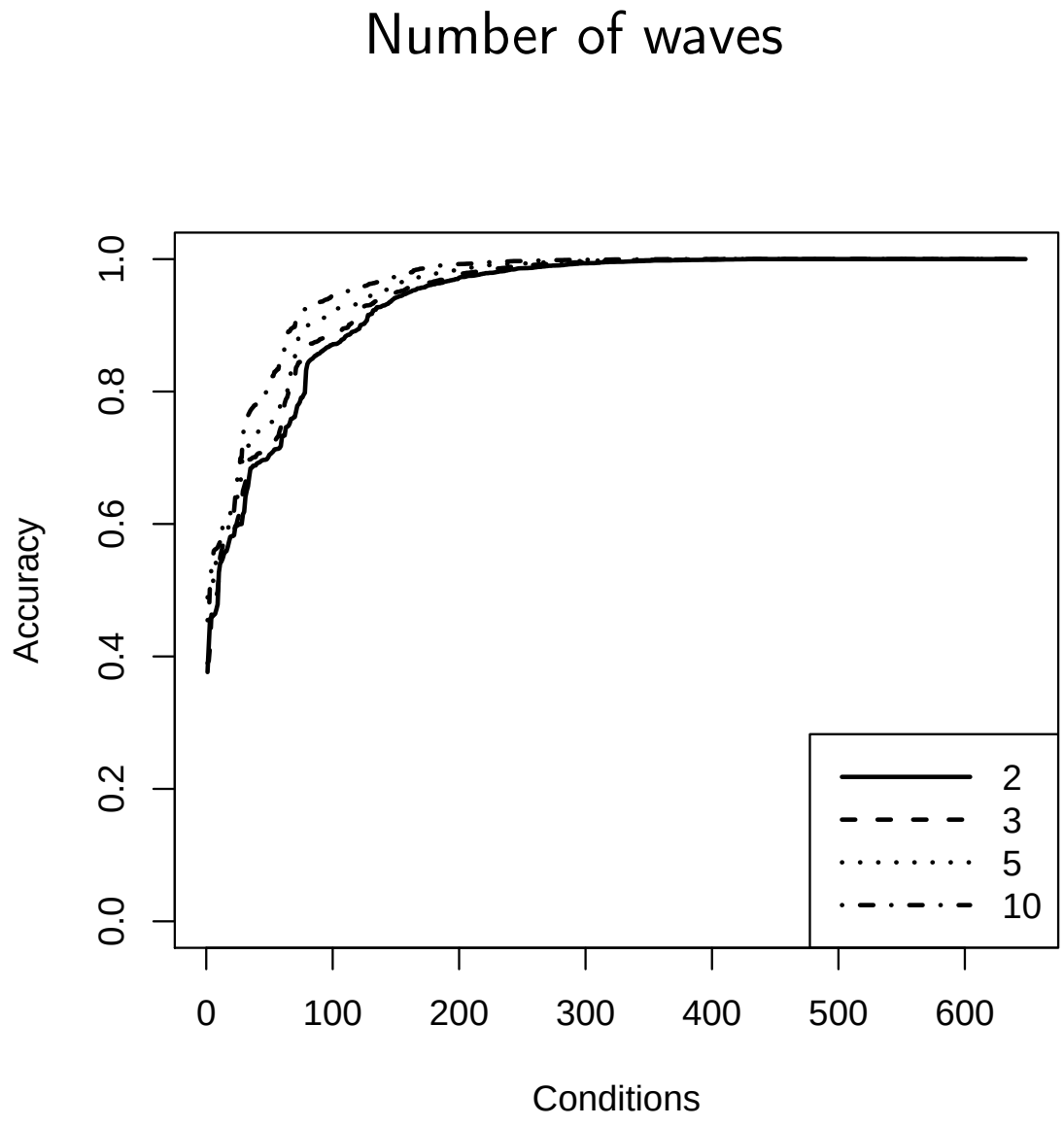
Influence of each factor



Influence of each factor



Influence of each factor

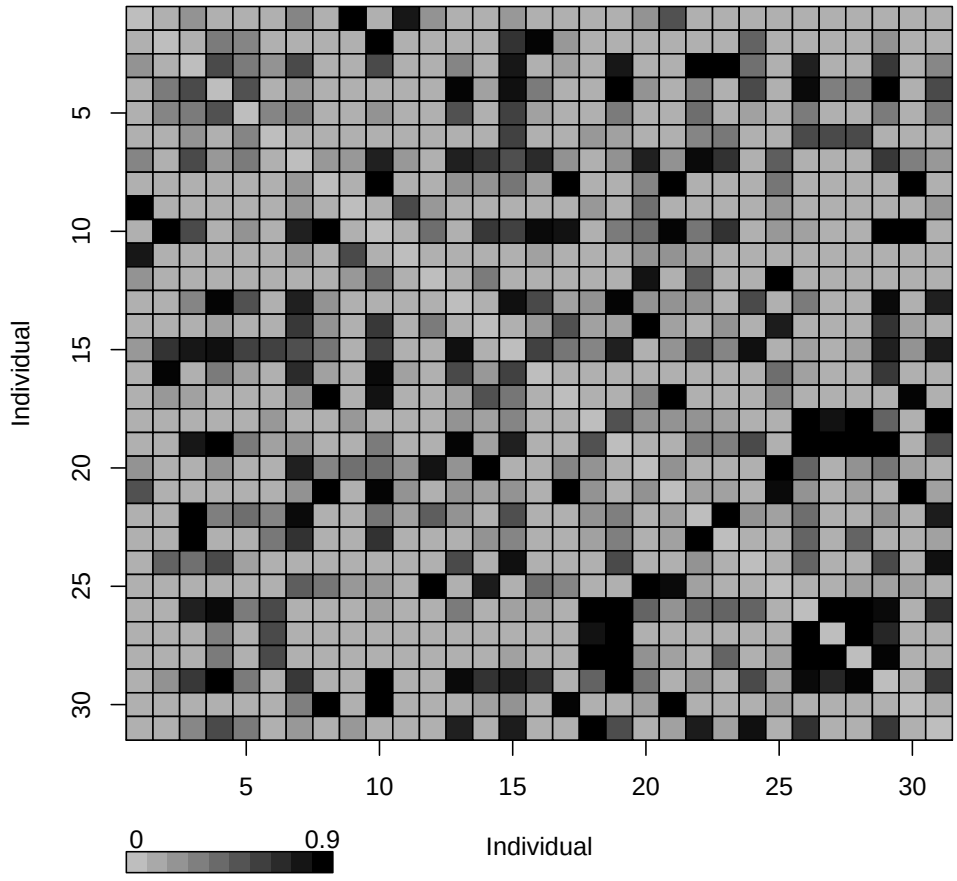


Real data analysis

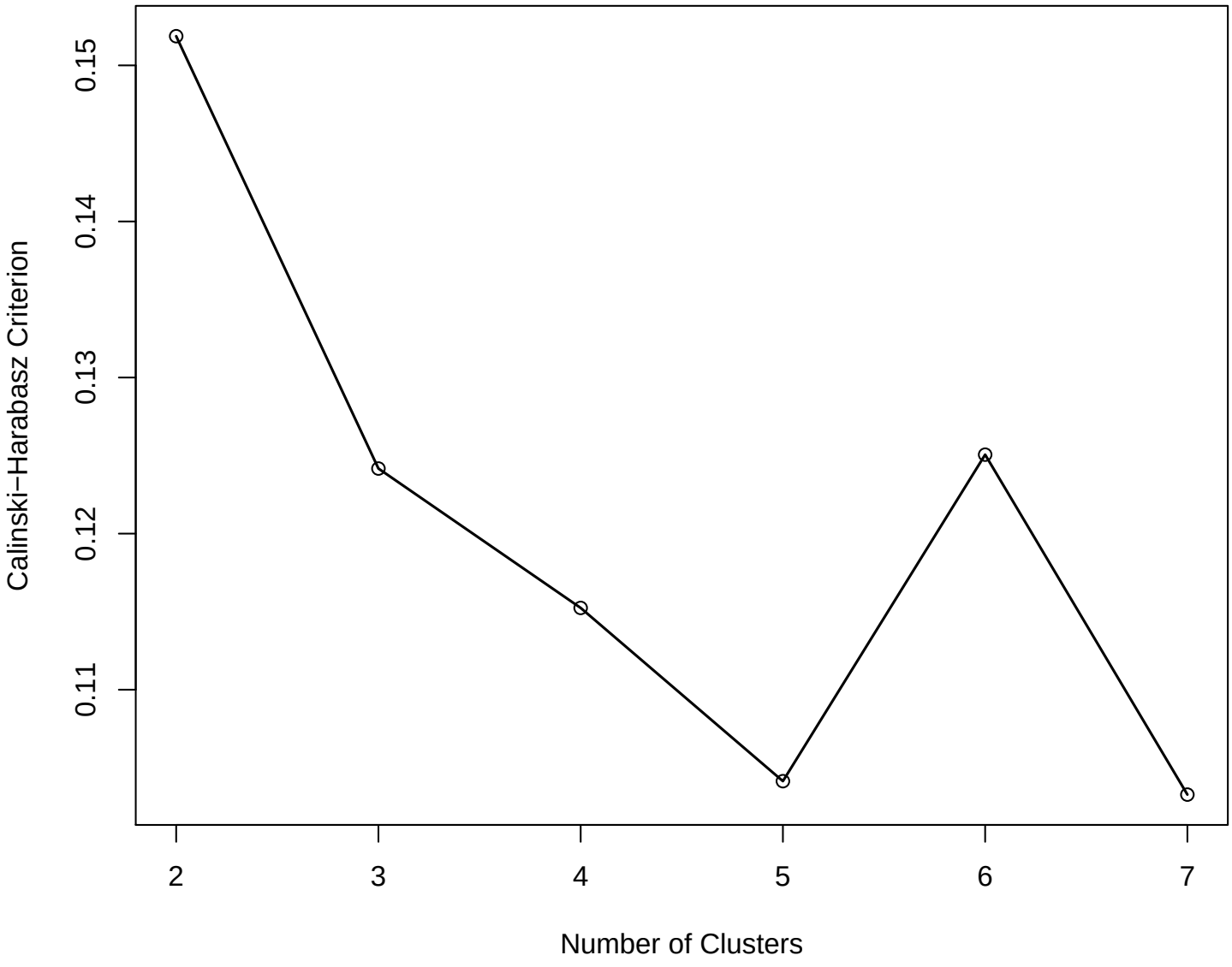
- The 7 waves of longitudinal network.
- Our new method
- Clustering based on the last wave of data
- Clustering based on the aggregated network

Strength matrix

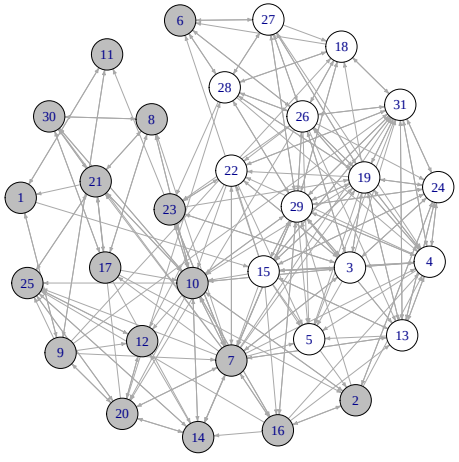
- The estimated tie strength ranges from 0.07 to 0.95.



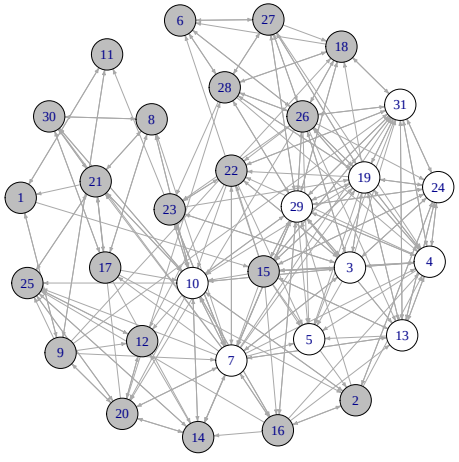
Determining the number of clusters



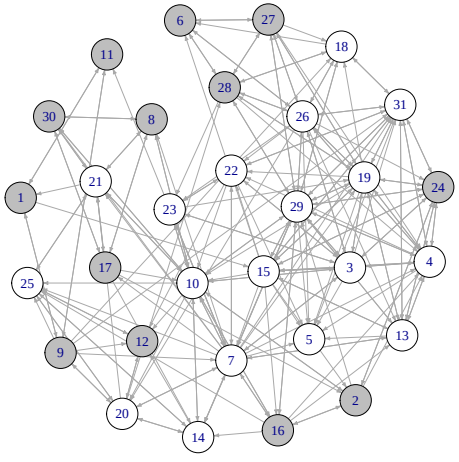
Comparison



(a) Our method



(b) Last wave



(c) Aggregated

Differences between the two clusters

			Estimate	Standard Error	Sig
Tie strength	C1	Density	-1.61	0.363	***
		Reciprocity	2.21	0.226	***
		Popularity	-0.22	0.191	
	C2	Density	1.26	0.545	***
		Reciprocity	2.27	0.312	***
		Popularity	-0.78	0.258	***
Last wave	C1	Density	2.06	1.185	
		Reciprocity	3.32	0.832	***
		Popularity	-1.33	0.649	*
	C2	Density	-0.42	0.340	
		Reciprocity	2.47	0.205	***
		Popularity	-0.66	0.198	***
All		Density	-0.33	0.200	
		Reciprocity	2.23	0.122	***
		Popularity	-0.38	0.088	***

Conclusions

- We proposed a new method for longitudinal social network clustering.
- 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.
- The real data analysis showed that our method can identify structural differences in the clusters.

Limitations and future directions

- We only evaluate one way to measure tie strength. Although it works for friendship relationship, other measures can be developed for different networks.
- When estimating tie strength, we did not take into account of 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.
- Clustering methods other than spectral clustering can be applied.
- Models other than the actor-based model can be used in longitudinal network generation.

Acknowledgments

- Institute of Education Sciences of U.S. Department of Education (R305D140037)
- National Science Foundation (1461355)
- ISLA at Notre Dame

Q & A

- For more information
  - Zhiyong Zhang (zzhang4@nd.edu)
  - Website: `http://bigdatalab.nd.edu`