5-2014

Persistent Community Detection in Dynamic Social Networks

Siyuan LIU
Carnegie Mellon University

Shuhui WANG
Institute of Computing Technology Chinese Academy of Sciences

Ramayya KRISHNAN
Carnegie Mellon University

Persistent Community Detection in Dynamic Social Networks

Siyuan Liu[1], Shuhui Wang[2], Ramayya Krishnan[1]

[1] Heinz College, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
{siyuan, rk2x}@cmu.edu
[2] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, 100190, China
wangshuhui@ict.ac.cn

Abstract. While community detection is an active area of research in social network analysis, little effort has been devoted to community detection using time-evolving social network data. We propose an algorithm, Persistent Community Detection (PCD), to identify those communities that exhibit persistent behavior over time, for usage in such settings. Our motivation is to distinguish between steady-state network activity, and impermanent behavior such as cascades caused by a noteworthy event. The results of extensive empirical experiments on real-life big social networks data show that our algorithm performs much better than a set of baseline methods, including two alternative models and the state-of-the-art.

Keywords: Community detection, persistent behavior, social networks.

1 Introduction

Much effort has been devoted to the development of community detection algorithms [7, 2, 3, 5], which can be used to identify clusters of nodes in social network data whose connections exhibit similar tendencies. Such clusterings may be intended for use as a predictive feature, or as a crude summary of network structure. In empirical studies, these algorithms often produce clusters that agree with general intuition about the network that is being studied, corresponding closely with known affiliations or genres held by the network nodes.

Here we develop a community detection model for *time-evolving* network data, and use this model to analyze a real-world call network. This data set is challenging to analyze, in part because of its large size (3.6 million users), and more importantly, because its structure also appears to change over time. To illustrate that the network is time-varying, Figure 1 shows a Q-Q plot of the degree distribution for calls made between 7-8AM and 8-9AM on a single day. The plot suggests that 7-8AM exhibits higher call density, and also a more heavy-tailed distribution (i.e., the largest values of the degree distribution grow more extreme). However, it is unclear to what extent the change in layout is due to time-varying structure (as opposed to being an artifact of the visualization process), and more importantly, how to quantify our observations.

In light of these concerns, we take a statistical model-based approach. Statistical modeling of dynamic network structure is challenging, and still nascent as a field of

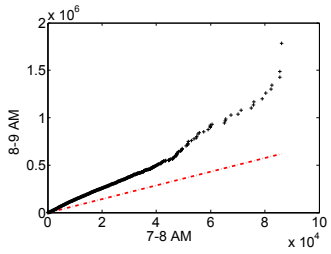


Fig. 1. Dynamics in mobile phone social networks on March 3^{rd}, 2008. Q-Q plot of call volumes from 7-8 AM and 8-9 AM.

research. Sociological theories for dynamic networks are not as well developed as for static networks, and hence less guidance is available for modeling. Perhaps due to this, existing time-varying models are typically designed to model many potential types of behavior [14]. Here we take a different approach. Our model is designed to detect only a single type of behavior; specifically, we find communities that exhibit *persistent* levels of communication over time. Our motivation is to distinguish between steady-state activity and impermanent behavior, such as cascades caused by a noteworthy event. We feel that persistence is the simplest type of dynamic behavior, making it a logical next step from the static setting.

Our contributions are summarized as follows.

1. We formally define a new network structure, *persistent community*, which exhibits persistent behavior/ structure over time.
2. We propose a novel algorithm to detect persistent community by a time and degree corrected blockmodel. We also provide inference of the model.
3. We conduct extensive empirical experiments on real-life big social networks data. Interesting findings and discussions are provided.

The rest of the paper is organized as follows. Section 2 surveys the previous work on community detection in dynamic networks. Section 3 proposes our algorithm. The empirical experiment results are reported in Section 4. At last, we conclude our work and give the future research directions in Section 5.

2 Related Work

The literature on static community detection is very large. Various extensions to the basic community model have been proposed, such as overlapping or mixed community membership [1, 13], degree-corrections which allow for heterogeneity within communities [6, 16], and community detection from trajectories [12]. In particular, without degree-correction, maximum likelihood methods often group the nodes according to their degree (i.e. their number of neighbors) [6]. As such behavior is typically undesirable, we will also include degree-corrections in our model.

Recent attention has been paid to community detection in dynamic social networks. Existing approaches, such as [7, 8, 4], generally detect communities separately for each time slot, and then determine correspondences by various methods [15, 18, 17]. However, such approaches often result in community structures with high temporal variation [10, 9]. Our approach differs in that inference in performed jointly across time, so as to find communities with low temporal variation (excepting network-wide fluctuations in call volume, such as night/day and weekday/weekend cycles).

3 Methodology

Our model can be considered to be a time-varying version of degree-corrected blockmodel in [6], specialized to directed graphs, where the expected call volumes within each community are assumed to all follow a single network-wide trajectory over time.

3.1 Time and degree-corrected blockmodel

General model We assume a network of N nodes over T time periods, where nodes are free to leave and re-enter the network. Let $\mathcal{I}_t \subset \{1, \ldots, N\}, t = 1, \ldots, T$ denote the subset of nodes which are present in the network at time t. Let K denote the number of communities, which determines the model order. Let $A^{(t)} \in \mathbb{N}^{N \times N}$ denote a matrix of call counts at time $t = 1, \ldots, T$; i.e., for $i \neq j$, $A_{ij}^{(t)}$ denotes the number of calls from node i to node j at time t. Our model is that the elements of $A^{(1)}, \ldots, A^{(T)}$ are independent Poisson random variables, whose parameters are jointly parameterized (with explanation of all parameters to follow):

$$
\begin{aligned}
A_{ij}^{(t)} &\sim \text{Pois}\left(\lambda_{ij}^{(t)}\right) \\
\lambda_{ij}^{(t)} &= \begin{cases} \theta_i^{(t)} \phi_j^{(t)} \mu^{(t)} \omega_{g_i g_j}^{(t)} & \text{if } i \in \mathcal{I}_t, j \in \mathcal{I}_t \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

We see that the expected number of calls λ_{ij} for each dyad (i, j) is a function of parameters g, ω, θ, ϕ, and μ. We now describe each parameter, its allowable values, and its function:

1. The vector $g \in \{1, \ldots, K\}^N$ assigns each node to a latent community in $1, \ldots, K$.
2. The matrix $\omega^{(t)} \in \mathbb{R}^{K \times K}$ gives the expected total call volume between each community at time $t = 1, \ldots, T$. In other words, $\omega_{ab}^{(t)}$ is the expected call volume from community a to community b at time t. To enforce persistence, ω is restricted to satisfy the following constraint:

$$
\omega_{aa}^{(t)} = \omega_{aa}^{(t')} \qquad t, t' \in 1, \ldots, T,\ a \in 1, \ldots, K. \tag{1}
$$

   As a result, intra-community call volumes are modeled as being constant over time (up to the network-wide effect of μ, which we discuss shortly), while inter-community call volumes may follow arbitrary trajectories over time.

3. The vector $\theta^{(t)} \in [0,1]^N$ controls the out-degree for each node at time $t = 1, \ldots, T$. Nodes whose element in $\theta^{(t)}$ is high will have higher expected outgoing call volumes than whose with low values in $\theta^{(t)}$. This allows for heterogeneity within communities. For identifiability, θ is restricted to satisfy the following constraint:

$$\sum_{i \in \mathcal{I}_t, g_i = a} \theta_i^{(t)} = 1 \qquad t = 1, \ldots, T, \ a = 1, \ldots, K$$

   The effect of this constraint is that $\omega_{ab}^{(t)}$ determines the total number of calls from community a to community b, while θ determines the proportion of calls emanating from each node in community a.
4. The vector $\phi^{(t)} \in [0,1]^N$ controls the in-degree for each node at time $t = 1, \ldots, T$, but is otherwise analogous to θ. A similar constraint is also enforced:

$$\sum_{i \in \mathcal{I}_t, g_i = a} \phi_i^{(t)} = 1 \qquad t = 1, \ldots, T, \ a = 1, \ldots, K.$$

5. The scalar $\mu^{(t)} \in [0,1]$ for $t = 1, \ldots, T$, modifies the total network call volume as a function of t. This allows for network-wide trends to be modelled, such as day/night or weekday/weekend cycles. For identifiability, μ is restricted to satisfy the following constraint:

$$\sum_{t=1}^{T} \mu^{(t)} = 1.$$

   The effect of this constraint is that $T\omega_{aa}^{(1)}$ determines the total number of calls within community a, while μ determines the proportion of those calls occurring within each time slot.

Discussion While self-calls are disallowed in a phone network, we note that our model assigns nonzero probability to positive values of $A_{ii}^{(t)}$. This is a simplification which decouples estimation of $\theta^{(1:T)}, \phi^{(1:T)}, \mu^{(1:T)}$ and $\omega^{(1:T)}$, leading to analytically tractable expressions for the parameter estimates. Self-calls predicted under the model should be disregarded as a modeling artifact. As the number of predicted self-calls will be a vanishing fraction of the total call volume, the effect will be negligible.

In the data, there exist pairs of individuals with extremely high call volumes, exceeding an average of 10 calls to each other per day. Such pairs are very sparse in the data ($< 1\%$ of all dyads), and do not seem to conform to the idea of community-based calling behavior. It is unlikely that the Poisson-based community model will explain these dyads. As such, we have opted to treat these dyads as outliers, and remove them before estimating the model parameters. Our interpretation is that the data is best described by the community based model, plus a small set of dyads whose high call volumes distinguish them from the overall network.

We note that θ and ϕ involve large numbers of parameters, as they are allowed to vary over time. A simpler model, in which θ and ϕ are constant over time, was considered. However, formulas for parameter inference become significantly more complicated in this case, unless $\mathcal{I}_t$ is also constant over time. If $\mathcal{I}_t$ is constant over time, so that

nodes cannot enter and leave the network, then the equations to be presented in Section 3.2 may be used with only slight modification if θ and ϕ are held constant over time.

3.2 Inference

We will estimate g and $\{\theta^{(t)}, \phi^{(t)}, \mu^{(t)}, \omega^{(t)}\}_{t=1}^{T}$ by maximum likelihood. We show here that given g, the maxmizing values of the remaining parameters can be found analytically, so that the maximizing the likelihood consists of a search over all community assignments in $\{1, \ldots, K\}^N$. While this exact maximization is computationally intractable, heuristic methods seem to give good results in practice, and we use a greedy search method described in [6] with multiple restarts.

We now derive formula for the remaining parameter estimates given g. The joint distribution of $A^{(1:T)} \equiv (A^{(1)}, \ldots, A^{(T)})$ (or equivalently the likelihood) is given by the product of Poisson distributions

$$L(\theta, \phi, \omega, \mu, g; A^{(1:T)}) = \prod_{t=1}^{T} \prod_{i,j \in \mathcal{I}_t} \frac{\left(\theta_i^{(t)} \phi_j^{(t)} \mu^{(t)} \omega_{g_i g_j}^{(t)}\right)^{A_{ij}^{(t)}}}{A_{ij}!} \times \exp\left(-\theta_i^{(t)} \phi_j^{(t)} \mu^{(t)} \omega_{g_i g_j}^{(t)}\right).$$

This expression can be simplified using the following intermediate terms. Given g and $A^{(1:T)}$, for all i, j, t let:

$$d_{i\cdot}^{(t)} = \sum_{j \in \mathcal{I}_t} A_{ij}^{(t)} \qquad d_{\cdot i}^{(t)} = \sum_{j \in \mathcal{I}_t} A_{ji}^{(t)},$$
$$m_{ab}^{(t)} = \sum_{i,j \in \mathcal{I}_t} A_{ij}^{(t)} 1\{g_i = a, g_j = b\},$$
$$m_{aa}^{(\cdot)} = \sum_{t=1}^{T} m_{aa}^{(t)} \qquad m_{\cdot\cdot}^{(t)} = \sum_{a=1}^{K} m_{aa}^{(t)}.$$

In words, $d_{i\cdot}^{(t)}$ and $d_{\cdot i}^{(t)}$ are the out-degree and in-degree of node i at time t; $m_{ab}^{(t)}$ is the call volume between communities a and b at time t; $m_{aa}^{(\cdot)}$ is the total call volume within community a over all time, and $m_{\cdot\cdot}^{(t)}$ is the total intra-community call volume (versus inter-community call volume) at time t. Using these terms, the likelihood L can be written as

$$L(\theta, \phi, \omega, \mu, g; A^{(1:T)}) = \frac{1}{\prod_{t,i,j} A_{ij}!} \prod_{t=1}^{T} \prod_{i=1}^{N} \left(\left[\theta_i^{(t)}\right]^{d_{i\cdot}^{(t)}} \left[\phi_i^{(t)}\right]^{d_{\cdot i}^{(t)}}\right) \times \prod_{t=1}^{T} \prod_{a,b=1}^{K} \left(\left[\mu^{(t)} \omega_{ab}^{(t)}\right]^{m_{ab}^{(t)}} \exp\left(-\mu^{(t)} \omega_{ab}^{(t)}\right)\right),$$

where we have used the constraints that $\sum_{i\in\mathcal{I}_t,g_i=a}\theta_i^{(t)}$ and $\sum_{i\in\mathcal{I}_t,g_i=a}\phi_i^{(t)}=1$. The function $\ell\equiv\log L$ is given by

$$\begin{aligned}\ell(\theta,\phi,\mu,g)=&\sum_{t=1}^{T}\sum_{i=1}^{N}\left(d_{i\cdot}^{(t)}\log\theta_i^{(t)}+d_{\cdot i}^{(t)}\log\phi_i^{(t)}\right)\\&+\sum_{t=1}^{T}\sum_{a,b=1}^{K}\left(m_{ab}^{(t)}\log\left[\mu^{(t)}\omega_{ab}^{(t)}\right]-\mu^{(t)}\omega_{ab}^{(t)}\right).\end{aligned}$$

We observe that ℓ can be grouped into terms which can be separately maximized,

$$\begin{aligned}\ell(\theta,\phi,\mu,g)=&\sum_{t=1}^{T}\sum_{i=1}^{N}\left(d_{i\cdot}^{(t)}\log\theta_i^{(t)}+d_{\cdot i}^{(t)}\log\phi_i^{(t)}\right)\\&+\sum_{t=1}^{T}\sum_{a=1}^{K}\left(m_{aa}^{(t)}\log\omega_{aa}^{(t)}-\mu^{(t)}\omega_{aa}^{(t)}\right)\\&+\sum_{t=1}^{T}\sum_{a=1}^{K}m_{aa}^{(t)}\log\mu^{(t)}\\&+\sum_{t=1}^{K}\sum_{a\neq b}\left(m_{ab}^{(t)}\log\left[\mu^{(t)}\omega_{ab}^{(t)}\right]-\mu^{(t)}\omega_{ab}^{(t)}\right).\end{aligned}$$

For fixed g, the maximizing value of the other parameters $\{\theta^{(t)},\phi^{(t)},\mu^{(t)},\omega^{(t)}\}_{t=1}^{T}$ can be analytically shown to satisfy for $t=1,\ldots,T$:

$$\begin{aligned}\mu^{(t)}&=\frac{m_{\cdot\cdot}^{(t)}}{\sum_{\tau=1}^{T}m_{\cdot\cdot}^{(t)}}\\\omega_{ab}^{(t)}&=\frac{m_{ab}^{(t)}}{\mu^{(t)}}&a\neq b\\\omega_{aa}^{(t)}&=m_{aa}^{(\cdot)}&a=1,\ldots,K\\\theta_i^{(t)}&=\frac{d_{i\cdot}^{(t)}}{\sum_{i\in\mathcal{I}_t}d_{i\cdot}1\{g_i=a\}}&i=1,\ldots,N\\\phi_i^{(t)}&=\frac{d_{\cdot i}^{(t)}}{\sum_{i\in\mathcal{I}_t}d_{\cdot i}1\{g_i=a\}}&i=1,\ldots,N\end{aligned}$$

Subsitution of the optimal values yields a function of g:

$$\begin{aligned}\ell(g)=&\sum_{t=1}^{T}\sum_{a=1}^{K}\left[H\left(\{d_{i\cdot}^{(t)}\}_{i\in\mathcal{I}_t,g_i=a}\right)+H\left(\{d_{\cdot i}^{(t)}\}_{i\in\mathcal{I}_t,g_i=a}\right)\right]\\&+\sum_{t=1}^{T}\sum_{a\neq b}h(m_{ab}^{(t)})+\sum_{a=1}^{K}h(m_{aa}^{(\cdot)})+H\left(\{m_{\cdot\cdot}^{(t)}\}_{t=1}^{T}\right),\end{aligned}\tag{2}$$

where the mapping H is given by $H(\{x_i\}_{i=1}^k) = \sum_{i=1}^k x_i \log \frac{x_i}{\sum_{j=1}^k x_j}$ for $x \in \mathbb{R}_+^k$, and the mapping h is given by $h(x) = x \log x - x$ for $x \in \mathbb{R}_+$.

We estimate the model parameters by optimizing $\ell(g)$ over all group assignments $g \in \{1, \ldots, K\}^N$. While it is intractable to find a global maximum, a local maximum can be found using the method described in [6]. After multiple restarts, the highest scoring local optima was chosen for the parameter estimate.

4 Empirical experiment results

4.1 Description of data and fitting procedure

The call records correspond to all mobile calls involving a particular service provider, with origin and destination within a particular city region, in the year 2008. The city area is roughly 8700 km^2, is covered by 5120 base stations, and serves 3.6 million mobile phone users. The data set is roughly 1 TB in size (more than 10 billion phone call records). The data set was prepared by the service provider, for the purpose of data mining to improve service and marketing capabilities. It contains call metadata (such as phone number, date of call, instant location of call), linked with customer profile information.

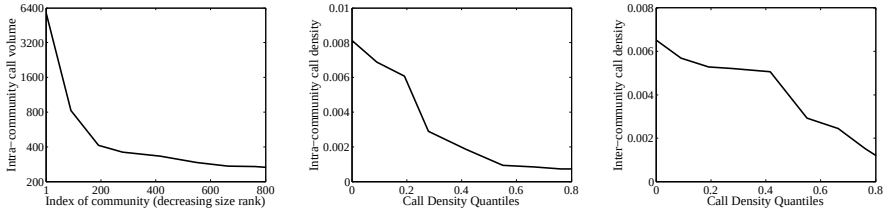
Using the call metadata, a set of directed adjacency matrices $\left(A^{(1)}, \ldots, A^{(365)}\right)$ was created, with nodes corresponding to customers (i.e., phone numbers), and edge weights corresponding to the number of calls between the sender and receiver on each day of the year. Community labels g were chosen to maximize ℓ as given by Eq.(2), using the algorithm described in Section 3.2, with the number of groups K chosen to be 800.

4.2 Out of sample prediction

To test the model, out of sample prediction was conducted, by randomly withholding 5% of the dyads from the fitting procedure. After fitting, the probability of connection according to the model (i.e., $P\left(A_{ij}^{(t)} > 0\right)$) was used to predict which of the withheld dyads had non-zero call volume. The model was highly predictive of the withheld dyads, with precision is 0.74 ± 0.05 and recall is 0.53 ± 0.05.

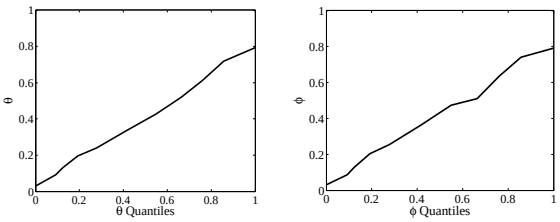
4.3 Description of model fit

To describe the fitted model, we give the following statistics. From the CDF giving the fraction of customers belong to the k largest communities, for $k = 1, \ldots, 800$, it shows that the majority of customers are concentrated in the 10 largest communities by the algorithm. Figure 2 (a) shows the fitted intra-community call densities, i.e., the call volume divided by the number of dyads in each community. The figure shows that the network is quite dense, with 30% of users in the largest community, in which every pair of members experienced an average of 0.008 phone calls over the course of a year. Figure 2 (b) plots the call densities (i.e., normalizing by the square of the community



(a) Intra-community call volume distribution (b) Intra-community call volume density distribution (c) Inter-community call volume density distribution

Fig. 2. (a) Call volumes for within-community communication, (b) Call densities for within-community communication, (c) Call densities for between-community communication



(a) θ quantiles. (b) ϕ quantiles.

Fig. 3. Fitted values of in-degree and out-degree corrections θ and ϕ, in ranked order.

size). Figure 2 (c) shows the fitted inter-community call volumes $\omega_{ab}^{(t)}$, for $a \neq b$ and $t = 1, \ldots, T$, as a quantile-plot. Based on Figure 2 (b) and Figure 2 (c), we see that the inter- and intra- community parameters follow different different distributions. Figure 3 (a) and 3 (b) shows the fitted degree corrections θ and ϕ as quantile plots. Figure 4 shows the fitted time-corrections $\mu^{(1)}, \ldots, \mu^{(T)}$. We note that larger values of μ occur on holidays and weekends.

To further understand the inferred communities, we compared the community labels g with groupings produced by various customer covariates included in the customer profile information:

- Age: the age of the customer, grouped by increments of six months.
- Gender: the gender of the customers.
- Workplace: the geographic region containing the registered workplace address of the customer. In our record, we have almost ten thousand workplace address and half a million customers provide this information.
- Residence: the geographic region containing the registered home address of the customer. In our record, we have almost ten thousand workplace address and half a million customers provide this information.
- Shopping mall: the shopping mall location. In our record, we have fifty shopping mall locations. On the other hand, based on the location information of each call, we are able to localize each customer, as reported in [11].

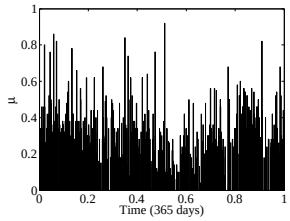


Fig. 4. Time-corrections $\mu^{(t)}$ as a function of time.

– Occupation: The occupational category reported by the customer. In our record, one hundred thousand customers have this information.

Table 1 (second column) reports the Jacaard similarity between the inferred community labels g and the customer covariates [3]. We find that the model has high similarity with Age, Workplace, and Residence, and Occupation, but not with Gender and Shopping Mall.

Table 1. Community Detection Evaluation on PCD

Covariate	PCD	A1	A2
Age	0.713	0.387	0.331
Gender	0.137	0.007	0.008
Workplace	0.728	0.411	0.527
Residence	0.617	0.208	0.317
Shopping Mall	0.21	0.087	0.012
Occupation	0.678	0.310	0.423

Figure 5 (a) shows the correlation coefficient for the intra-community and inter-community call volumes, for time lags varying from 10 to 40 weeks.

The figure shows that the observed intra-community call volumes are more persistent over time compared to the inter-community call volumes, suggesting that the method is successful in finding communities with persistent intra-community call volumes.

4.4 Work v.s. leisure groupings

To differentiate work and leisure interactions, the data was separated into weekdays and weekends, and then g was fit separately by Eq.(2) on the two scenarios. As shown in Table 2, we found that the weekday groupings corresponded more closely to (place of

[3] $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, where A and B are two label sets.

employment, or some other covariate), which the weekend groups were more closely aligned with family relationships (which are recorded in the data set).

We also notice that the usage of persistence constraints had a larger effect in the weekend groups; this suggests that the social/weekend groups are less visible in the data (i.e., a "weaker signal"), causing the model regularization to have greater effect.

Table 2. Weekday grouping v.s. weekend grouping (Jacaard Similarity)

Covariate	Weekday	Weekend
Age	0.731	0.702
Gender	0.152	0.120
Workplace	0.801	0.568
Residence	0.578	0.817
Shopping Mall	0.124	0.453
Occupation	0.831	0.542

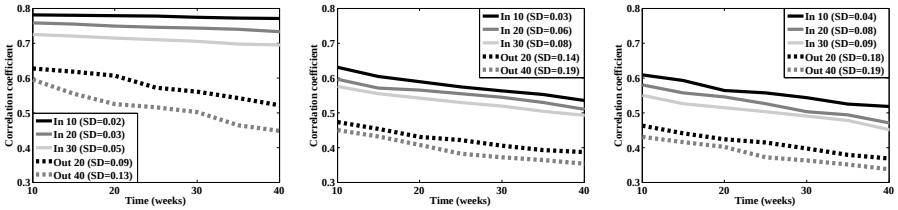
4.5 Comparison with other community detection algorithms

The results of our method were compared against several other algorithms. Specifically:

- A1: The persistence constraint $\omega_{aa}^{(t)} = \omega_{aa}^{(t')}$ given by Eq. 1 is removed when fitting the model. As a result, the intra-community expected call volumes are no longer constrained to follow any particular trajectory over time.
- A2: A static degree-corrected blockmodel, as described in [6], is fit to the static matrix $A = \sum_{t=1}^{T} A^{(t)}$.
- DSBM: A bayesian approach [17] for detecting communities in dynamic social networks.

Under the algorithm A1, as shown in Table 1 (third column), the groupings differed significantly compared to those found by our proposed method, and did not correspond as well to observed covariates, as described in Table 1 (second column). Figure 5 (b) shows the average correlation coefficients for the intra- and inter-community call volumes. We observe that the coefficients are lower compared to our proposed algorithm, suggesting that the call volumes are less persistent over time.

Under the algorithm A2, similar findings resulted, as shown in Table 1 (fourth column) and Figure 5 (c). It is interesting that the similarity results of A1 and A2 are different and can be interpreted by the methods we use. For A1, we release the persistence constraint for intra-community connection, while for A2, we use a static model. For Age and Gender, A1 and A2 give similar results, but for Workplace, Residence and Occupation, A2 can give much better similarity result than A1, while for Shopping Mall, A1 is better. It means in working places and resident locations, people prefer to make connections within communities, while in shopping mall locations, the dynamics of social connection is much stronger and impacted by time of day and day of week. We can further interpret the result as that human social behavior is not only impacted by real life behaviors, but also the time of day and day of week.



(a) Persistence evaluation of our PCD algorithm.

(b) Persistence evaluation of the alternative model 1.

(c) Persistence evaluation of the alternative model 2.

Fig. 5. Persistence evaluation. Within-group call volumes are persistent over time, up to network-wide trends. In contrast, the call volumes in the off-diagonal plots are much lower, and do not follow network-wide trends, suggesting that communication between groups was more sporadic. The persistence of the communities detected by alternative model 1 and 2 are not good as the result of PCD.

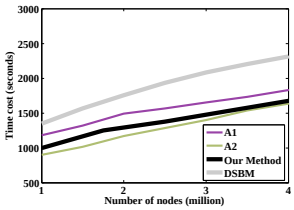


Fig. 6. Efficiency evaluation of different algorithms. The running time cost of our algorithm (PCD) is much lower than two other baseline algorithms.

4.6 Computational runtime

In Figure 6, we report computational runtimes for the different algorithms. The results show that our method runs much faster than two other baseline methods. For a fixed number of random restarts, the runtime scales nearly linearly for the graph sizes considered here. All algorithms were conducted on a standard server (Linux), with four Intel Core Quad CPUs, Q9550 2.83 GHz and 32 GB main memory.

5 Conclusion and future work

In this paper, we studied an interesting but challenging problem, persistent community detection in evolving social graphs. Extensive empirical experiment results show that our proposed method performs much better than a set of baseline methods, in merits of persistence in time series analysis, consistency in social graph structure and efficiency in algorithm running time cost.

In the future, we are going to apply our method to online social networks (e.g., Facebook and Twitter), and then we would like to compare the persistent community in mobile social networks with the persistent community in online social networks.

6 Acknowledgments

This research was supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA) and the Pinnacle Lab at Singapore Management University. Shuhui Wang was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, and National Natural Science Foundation of China: 61303160. The authors also thank David Choi for valuable discussions and support regarding this work.

References

1. A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A tensor spectral approach to learning mixed membership community models. *CoRR*, pages –1–1, 2013.
2. N. Barbieri, F. Bonchi, and G. Manco. Cascade-based community detection. In *WSDM'13*, pages 33–42, 2013.
3. R. J. D'Amore. Expertise community detection. In *SIGIR'04*, pages 498–499, 2004.
4. O. V. Drugan, T. Plagemann, and E. Munthe-Kaas. Detecting communities in sparse manets. *IEEE/ACM Trans. Netw.*, pages 1434–1447, 2011.
5. S. Fortunato. Community detection in graphs. *CoRR*, pages –1–1, 2009.
6. B. Karrer and M. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.
7. J. Leskovec, K. J. Lang, and M. W. Mahoney. Empirical comparison of algorithms for network community detection. In *WWW'10*, pages 631–640, 2010.
8. W. Lin, X. Kong, P. S. Yu, Q. Wu, Y. Jia, and C. Li. Community detection in incomplete information networks. In *Proc. of WWW 2012*.
9. Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. In *WWW'08*, pages 685–694, 2008.
10. Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Analyzing communities and their evolutions in dynamic social networks. *TKDD*, pages –1–1, 2009.
11. S. Liu, L. Kang, L. Chen, and L. M. Ni. Distributed incomplete pattern matching via a novel weighted bloom filter. In *ICDCS'12*, pages 122–131, 2012.
12. S. Liu, S. Wang, K. Jeyarajah, A. Misra, and R. Krishnan. TODMIS: Mining communities from trajectories. In *ACM CIKM*, 2013.
13. N. P. Nguyen, T. N. Dinh, S. Tokala, and M. T. Thai. Overlapping communities in dynamic networks: their detection and mobile applications. In *MOBICOM'11*, pages 85–96, 2011.
14. B. Skyrms and R. Pemantle. A dynamic model of social network formation. *Proceedings of the National Academy of Sciences of the United States of America*, 97(16):9340–9346, 2000.
15. L. Tang, X. Wang, and H. Liu. Community detection via heterogeneous interaction analysis. *Data Min. Knowl. Discov.*, pages 1–33, 2012.
16. X. Yan, J. E. Jensen, F. Krzakala, C. Moore, C. R. Shalizi, L. Zdeborov, P. Zhang, and Y. Zhu. Model selection for degree-corrected block models. *CoRR*, pages –1–1, 2012.
17. T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin. Detecting communities and their evolutions in dynamic social networks - a bayesian approach. *Machine Learning*, pages 157–189, 2011.
18. Y. Zhang, J. Wang, Y. Wang, and L. Zhou. Parallel community detection on large networks with propinquity dynamics. In *KDD'09*, pages 997–1006, 2009.