Clustering in a growing network: a dynamic extension of stochastic block model*

Zitong Zhang[†,‡,??], **Shizhe Chen**[§,??]

Address of the First and Second authors
Usually a few lines long
e-mail: zztzhang@ucdavis.edu

Abstract: We propose a dynamic extension of the stochastic block model for growing networks, where isolated nodes build connection with each other and develop into a well-organized network. Our model aims at analyzing typical roles of nodes in the growing procedure, and incorporating the time delay of edges caused by the various active time of nodes. We develop an algorithm based on k-means and shape invariant method. We demonstrate the performance of our algorithm on simulation experiments.

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35.
Keywords and phrases: sample, LaTeX 2$_\varepsilon$.

1. Introduction

Dynamic networks emerge in many area, such as the neuronal network in the brain during disease[] or learning tasks[], the social network in a time period[], etc. A lot of work has been done in order to study the dynamic networks, but most of these research focus on analyzing the dynamics in a well-developed network. In this paper, we concentrate on the growing networks where a bunch of isolated nodes are developed into a functional network. To the best of our knowledge, there is no existing study about the growing networks, partly because of the lack of data. Fortunately, the data collected by Wan et al. [9] provides us the possibility to pursue this study.

In this paper, we focus on the neural data provided by [9]. Since the neuronal network is complicated, we try to break it down and propose to model it by identifying the typical roles of individual neurons. As supported by [9], neurons in a growing network play different roles in terms of active time and connecting patterns. The connecting pattern of a neuron can be described as the occurrence time of edges that include this neuron. Neurons with the same roles perform similar activities and thus have similar connecting patterns. So the roles of

*Footnote to the title with the "thankstext" command.
†Some comment
‡First supporter of the project
§Second supporter of the project

neurons are the basic units in a complex network, and learning these roles can help us understand the developing procedure of growing networks.

However, being able to identify the typical roles in a single growing network is infeasible because the growing network data is transient — the time of the growing trajectory cannot go to infinity. Due to this limit, we need our method to be able to borrow information from other networks. This is not trivial, because the neurons in different networks are not one-to-one mapped, and so the roles identified from different subjects are not transferable, . This constrains our ability to study the common features across subjects. For this reason, we propose to use the stochastic block model as it allows to combine multiple networks and hence resolve the above problems. [SBM is.... It has been applied to ...]

There are, nevertheless, some unique properties of the data that are beyond the scope of the stochastic block model. First, the connection between two neurons is measured over time. Second, the connecting pattern between two neurons are determined not only by their roles but also by their active time, which varies from node to node. Third, the connection is also effected by the spatial distance between neurons — connection cannot occur if two neurons are too away from each other. To incorporate such uniqueness, we propose a generalized stochastic block model in this paper. [Note that these properties also appear in other problems, e.g. venmo...]

Related work

The stochastic block model is first proposed by Holland, Blackmond and Leinhardt [4]. It has many dynamic extensions, Yang et al. [12], Xu and Hero [11], Matias and Miele [5], Xu [10] use the Markov chain to model the time-varying connecting probabilities and/or the clustering matrix. EM algorithm or iterative optimization algorithm is commonly used for inference.

Matias, Rebafka and Villers [6] adapt the stochastic block model to the context of recurrent interaction events in continuous time, where the recurrent events are modeled by Poisson processes with intensities determined by the nodes' group memberships. The maximum likelihood estimator is proposed, but no theoretical analysis is available in the paper.

Optimal rate of convergence is also studied. Gao, Lu and Zhou [3] provides an optimal rate under the mean squared error for the stochastic block model. Pensky [7] derives a penalized least square estimator in a dynamic network setting, and shows that the estimator satisfies an oracle inequality and attains the minimax lower bound for the risk.

Contribution

In this paper, we propose a method for analyzing the growing networks. Our method is able to identify the roles of individual nodes and the connecting patterns. We adapt the stochastic block model to the growing networks context by generalizing the connecting probabilities to intensities of point processes.

In addition, we incorporate the time delay of each node so that our model is able to handle the network where nodes become active over time. We derive a least square estimator and show that the estimator converges [in a certain rate]. Finally, an algorithm combining the k-means method and the shape invariant method is proposed for estimation.

Future work

Future working directions include (but not limited to) (i) identifying clusters with similar connecting pattern but different active time phase or different vertex degree, (ii) incorporating the movement of nodes, (iii) seeking for a convex relaxation method that convexify over both clustering matrix and time lags (convex relaxation can also be adapted to solve the penalized least square problem in Pensky [7]), (iv) try other clustering methods.

Organization

The rest of this paper is organized as follows. In Section 2, we review the stochastic block model and introduce the proposed dynamic generalization of the stochastic block model. We introduce the least square estimator and the estimation algorithm in Section 3. Theoretical results are provided in Section 4. Section 5 shows the numerical experiments.

2. Model

2.1. Stochastic block model

A set of n nodes $\Gamma = \{1, \cdots, n\}$ is partitioned into k clusters $\Gamma_1, \cdots, \Gamma_k$. The cluster of node i is represented by $z_i \in \{1, \cdots, k\}$, and the vector of clusters is $\mathbf{z} = (z_i)_{i=1}^n$. Define the adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$ where $A_{i,j} = 1$ if an edge is observed between node i and node j and $A_{i,j} = 0$ otherwise. We set $A_{i,i} \equiv 0$ for any $i = 1, \cdots, n$, and assume that $A_{i,j}$'s are conditionally independent given the cluster vector $\mathbf{z}$:

$$A_{i,j}|z_i = q, z_j = l \overset{ind}{\sim} \text{Bernoulli}(C_{q,l}), \qquad i \neq j,$$

where $\mathbf{C} \in [0,1]^{k\times k}$ denotes the connecting probability matrix.

2.2. Dynamic generalization of the stochastic block model

In a growing network where the edges appear over time, the edge between a pair of nodes i and j can be represented by $N_{i,j}(\cdot)$ with intensify function

$$\lambda_{i,j}(t) = \lambda_{z_i,z_j}(t), \qquad t \in [0,T], \quad i,j = 1, \cdots, n,$$

where $[0,T]$ is overall time period, $\lambda_{z_i,z_j}(\cdot)$ is the connecting intensity function between cluster z_i and z_j. Similar to the stochastic block model, we set $\lambda_{i,i}(\cdot) \equiv 0$ for $i = 1, \cdots, n$.

In practice, there are usually more restrictions to the network. Motivated by the neuronal network (see [9] for detail) where neurons become active in different time and only connect with neighbor neurons, we incorporate the time delay and spatial location of each node and propose the following model:

$$\lambda_{i,j}(t) = \lambda_{z_i,z_j}(t-\tau_{i,j}) \cdot \mathbf{1}_{\{d_{i,j} \le d^*\}}, \qquad t \in [0,T], \quad i,j = 1, \cdots, n,$$

where T and λ_{z_i,z_j} is defined as before, $\tau_{i,j}$ is the time delay caused by both node i and node j, $d_{i,j}$ is the spatial distance between node i and j, and the node i and j can be connected only if $d_{i,j} \le d^*$.
For the convenience of estimation, we make the following assumptions.

Assumption 1. *$\tau_{i,j}$ only depends on node i, that is, $\tau_{i,j} = \tau_i$ for all $j \neq i, i = 1, \cdots, n$.*

With Assumption 1, we may consider the integrated point process $N_i(\cdot) := \sum_{j \neq i} N_{i,j}(\cdot)$ and write its intensity function as

$$\begin{aligned}\lambda_{N_i}(t) &= \sum_{l=1}^{k} \left(\lambda_{z_i,l}(t-\tau_i) \cdot \sum_{j \in \Gamma_l, j \neq i} \mathbf{1}_{\{d_{i,j} \le d^*\}} \right) \\ &=: \sum_{l=1}^{k} \lambda_{z_i,l}(t-\tau_i) \cdot w_{i,l}.\end{aligned}$$

Here $w_{i,l}$ is the number of nodes from cluster l that is in the neighborhood of node i.
We also assume that the nodes are distributed uniformly in the sense that $w_{i,l}$ and $w_{j,l}$ are identically distributed for any nodes i, j from the same cluster. More formally, we have the following assumption.

Assumption 2. *For any $l = 1, \cdots, k$, $\{w_{i,l}\}_{i \in \Gamma_l}$ are i.i.d. with mean $\bar{w}_{z_i,l}$ and variance $\sigma^2 < \infty$.*

By Assumption 2, $\lambda_{N_i}(t+\tau_i) \overset{d}{=} \lambda_{N_j}(t+\tau_j)$ for node i, j with $z_i = z_j$. And hence we can define the mean intensify function for each group $\lambda_l(t) \overset{\triangle}{=} \mathbb{E}\lambda_{N_i}(t+\tau_i), i \in \Gamma_l$. We aim at estimating the clusters $\mathbf{z}$, the mean intensity functions $\{\lambda_l(\cdot)\}_{l=1}^k$, and the time delays $\{\tau_i\}_{i=1}^n$.

3. Method

3.1. k-means objective function

In what follows, we review the k-means objective function in the Euclidean space, and introduce the objective function in our case where the samples are realizations of point processes.

k-means in $\mathbb{R}^d$

Let $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ be an i.i.d. sample from distribution function F. Denote by F_n the empirical distribution function. The k-means problem is to solve

$$A_n := \underset{A:A\subset\mathbb{R}^d,|A|=k}{\arg\min} W_n(A, F_n) = \underset{A:A\subset\mathbb{R}^d,|A|=k}{\arg\min} \int \min_{a\in A} \|\mathbf{x}_i - a\|^2 \mathrm{d}F_n.$$

There are some theoretical guarantees of k-means. Pollard [8] shows that for a given k, $A_n \to \bar{A}$ almost surely, where $\bar{A} = \arg\min_A W(A, F)$ denotes the optimal population cluster centers. Bachem et al. [1] provides a uniform bound with a rate of $\mathcal{O}\left(n^{-1/2}\right)$ for the deviation between the empirical loss and the expected loss. The bound is uniform in the sense that it holds for any set of k cluster centers.

Note that $\bar{A}$ is a biased estimator of the true cluster centers (when they are well-defined). For example, if $k = 2$ and $F(x) = \frac{1}{2}\Phi(x; \mu_1, \sigma_1^2) + \frac{1}{2}\Phi(x; \mu_2, \sigma_2^2)$ is a mixture Gaussian distribution, and denote $X_1 \sim N(\mu_1, \sigma_1^2)$, then $\bar{A} = \{a_1, a_2\}$ where $a_1 = \mathbb{E}[X_1 \mathbf{1}_{X_1 \le (\mu_1+\mu_2)/2}] \neq \mu_1$ (and a similar expression for a_2).

This problem can be re-formulated as following

$$\min_{\{\Gamma_l\}_{l=1}^k} \frac{1}{n} \sum_{l=1}^{k} \sum_{i\in\Gamma_l} \|\mathbf{x}_i - \mathbf{c}_l\|^2, \tag{3.1}$$

where $\{\Gamma_l\}_{l=1}^k$ represent the clusters and form a partition of $\Gamma = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, $\mathbf{c}_l = 1/|\Gamma_l| \cdot \sum_{\mathbf{x}_i\in\Gamma_l} \mathbf{x}_i$ is the sample center of the l-th cluster. For simplicity, we denote $\mathbf{x}_i \in \Gamma_l$ by $i \in \Gamma_l$ henceforth. We now extend this objective function to the context of point process.

k-means in point processes

By Assumption 1 and 2, $\lambda_{N_i}(t+\tau_i) \overset{d}{=} \lambda_{N_j}(t+\tau_j)$ for any i, j satisfying $z_i = z_j$. Thus $F_{N_i}(t+\tau_i) \overset{d}{=} F_{N_j}(t+\tau_j)$ for any i, j satisfying $z_i = z_j$, where $F_{N_i}(t) := \int_0^t \lambda_{N_i}(s)\mathrm{d}s \big/ \int_0^T \lambda_{N_i}(s)\mathrm{d}s$. So the k-means problem is to solve

$$\min_{\{\Gamma_l\}_{l=1}^k} \frac{1}{n} \sum_{l=1}^{k} \left(\min_{\{\tau_i\}_{i\in\Gamma_l}, F_l} \sum_{i\in\Gamma_l} \|\tilde{F}_{N_i}(\cdot+\tau_i) - F_l(\cdot)\|_2^2 \right), \tag{3.2}$$

where $\tilde{F}_{N_i}(t) := 1/N_i([0,T]) \cdot \sum_{j=1}^{N_i([0,T])} \mathbf{1}_{\{t_{N_i,j}\le t\}}$ is the empirical distribution function of the occurrence time of edges, $t_{N_i,j}$ is the occurrence time of the j-th edge of $N_i(\cdot)$, and $F_l(t) = \mathbb{E}[F_{N_i}(t+\tau_i)](\forall i \in \Gamma_l)$ is the expected cumulative distribution function of the l-th cluster.

3.2. Algorithm

The initialization method and the choice of the number of clusters k will be discussed later, for now we assume the initialization and k are given. To solve the problem (3.2), we iterate between two steps until convergence:

- Re-cluster step: update the clustering $\left\{\hat{\Gamma}_l\right\}_{l=1}^{k}$ based on the distance $d(\tilde{F}_{N_i}, \hat{F}_l)$ defined as

$$d(\tilde{F}_{N_i}, \hat{F}_l) = \min\Big\{ \inf_{\tau\in[0,T]} \left(\int_{-T}^{T} \left|S_\tau \circ \tilde{F}^*_{N_i}(t) - \hat{F}^*_l(t)\right|^2 \mathrm{d}t\right)^{1/2}, \\ \inf_{\tau\in[0,T]} \left(\int_{-T}^{T} \left|\tilde{F}^*_{N_i}(t) - S_\tau \circ \hat{F}^*_l(t)\right|^2 \mathrm{d}t\right)^{1/2} \Big\},$$

  where

$$S_\tau \circ \tilde{F}^*_{N_i}(t) = \begin{cases} 0, & t\in[-T,-\tau) \\ \tilde{F}_{N_i}(t+\tau), & t\in[-\tau, T-\tau) \\ 1, & t\in[T-\tau, T] \end{cases}, \hat{F}^*_l(t) = \begin{cases} 0, & t\in[-T,0) \\ \hat{F}_l(t), & t\in[0,T] \end{cases}. \tag{3.3}$$

- Re-center step: update the expected cumulative distribution functions $\{\hat{F}_l\}_{l=1}^{k}$ using the method in Bigot and Gendre [2].

In the re-cluster step, the distance $d(\tilde{F}_{N_i}, \hat{F}_l)$ for each pair of node i and cluster l is evaluated by solving the problem

$$\begin{aligned} \hat{n}_{i,l} &= \underset{|n|\le(N-1)/2}{\arg\min} \sum_{0<|j|\le(N-1)/2} \left|\theta_{i,j} e^{\mathrm{i}2\pi jn/N} + \frac{e^{\mathrm{i}2\pi jn/N}-1}{1-e^{\mathrm{i}2\pi j/N}} - \gamma_{l,j}\right|^2 + |\theta_0 + n - \gamma_0|^2 \\ &= \underset{|n|\le(N-1)/2}{\arg\min} \sum_{0<|j|\le(N-1)/2} \left|\left(\theta_{i,j} + \frac{1}{1-e^{\mathrm{i}2\pi j/N}}\right) e^{\mathrm{i}2\pi jn/N} - \left(\gamma_{l,j} + \frac{1}{1-e^{\mathrm{i}2\pi j/N}}\right)\right|^2 + \\ &\qquad |\theta_0 + n - \gamma_0|^2 \\ &\triangleq \underset{|n|\le(N-1)/2}{\arg\min} \sum_{0<|j|\le(N-1)/2} \left|\theta'_{i,j} e^{\mathrm{i}2\pi jn/N} - \gamma'_{l,j}\right|^2 + |\theta_0 + n - \gamma_0|^2, \end{aligned} \tag{3.4}$$

where $n = N\cdot\tau/2T$, $\theta_{i,j}$ and $\gamma_{l,j}$, $j = -(N-1)/2, \cdots, (N-1)/2$, are the discrete Fourier coefficients of $\tilde{F}^*_{N_i}$ and $\hat{F}^*_l$, N is the length of discretized version of $\tilde{F}^*_{N_i}$ and $\hat{F}^*_l$.

Gradient descent can be used to solve the problem. The gradient is shown below:

$$\nabla_{n_i} = \frac{4\pi}{N} \cdot \sum_{0<|j|\le(N-1)/2} j \cdot \mathrm{Im}\left(\theta'_{i,j}\overline{\gamma'_{i,j}} e^{\mathrm{i}2\pi n_i j/N}\right) + 2n_0 + 2\mathrm{Re}\left(\theta_0 - \gamma_0\right). \tag{3.5}$$

The learning rate was set as 0.01, the initialization of n_i was set to be the last estimated $\hat{n}_i$.

In the re-center step, let $\{\theta_{i,j}\}_{j\in\mathbb{Z}}$ and $S_\tau \circ \tilde{F}^*_{N_i}(t)$ be defined the same as above. The Fourier coefficients of $S_\tau \circ \tilde{F}^*_{N_i}(t)$ are

$$\theta_{i,j}e^{\mathrm{i}2\pi jn/N} + \frac{1}{1-e^{\mathrm{i}2\pi j/N}}\left(e^{\mathrm{i}2\pi jn/N}-1\right) \triangleq \theta'_{i,j}e^{\mathrm{i}2\pi jn/N} - C.$$

Then $\{\hat{\tau}_i\}_{i=1}^n = \{\hat{n}_i \cdot 2T/N\}_{i=1}^n$ can be obtained by

$$\{\hat{n}_i\}_{i\in\Gamma_l} = \underset{\min_{i\in\Gamma_l}\{n_i\}=0}{\arg\min} \frac{1}{|\Gamma_l|}\sum_{i\in\Gamma_l}\sum_{j=1}^{N}\left|\theta'_{i,j}e^{\mathrm{i}2\pi jn_i/N} - \frac{1}{|\Gamma_l|}\sum_{i'\in\Gamma_l}\theta'_{i',j}e^{\mathrm{i}2\pi jn_{i'}/N}\right|^2.$$

Using gradient descent over all $\{n_i\}_{i\in\Gamma_l}$ is expensive, thus we adopt the following two-step estimation procedure:

- Estimate the mean distribution function $\hat{F}^*_l(t)$.
- Estimate $\{\hat{n}_i\}_{i\in\Gamma_l}$ by aligning each $\tilde{F}^*_{N_i}(t)$ with $\hat{F}^*_l(t)$.

The initialization of $\hat{F}^*_l(t)$ can be any randomly chosen $\tilde{F}^*_{N_i}(t)$.

Finally, the mean distribution function is estimated as the average of shifted empirical distribution functions.

Initialization

The k-means++ method is applied for choosing initial mean curves.

Choosing k

4. Theory

Assume $\mathbf{F} = (F_i)_{i=1}^n$ is from the parameter space

$$\mathcal{F}_k = XXX.$$

Theorem 4.1. *For any constant $C' > 0$, there is a constant $C > 0$ only depending on C', such that*

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{F}_i - F_i\right\|^2 \le C\left(XXX\right),$$

with probability at least $1 - \exp\left(-C'XXX\right)$, uniformly over $\mathbf{F} \in \mathcal{F}_k$.

Proof. This is a sketch of proof and is based on the proof in [3].

We denote the true value by $\theta_i^* = F^*_{Z_i^*}(\cdot - \tau_i^*)$. For the estimated $\hat{z}$, define $\tilde{\theta} = \underset{\theta \in \Theta_k(\hat{z})}{\arg\min} \|\theta^* - \theta\|^2$. By the definition of the estimator, we have

$$L(\hat{F}, \hat{Z}, \hat{\tau}) \leq L(F^*, Z^*, \tau^*),$$

which can be rewritten as

$$\|\hat{\theta} - F^{obs}\|^2 \leq \|\theta^* - F^{obs}\|^2. \tag{4.1}$$

The left-hand side of (4.1) can be decomposed as

$$\|\hat{\theta} - \theta^*\|^2 + 2\langle \hat{\theta} - \theta^*, \theta^* - F^{obs} \rangle + \|\theta^* - F^{obs}\|^2. \tag{4.2}$$

Combining (4.1) and (4.2), we have

$$\|\hat{\theta} - \theta^*\|^2 \leq 2\langle \hat{\theta} - \theta^*, F^{obs} - \theta^* \rangle. \tag{4.3}$$

The right-hand side of (4.3) can be bounded as

$$\begin{aligned}
\langle \hat{\theta} - \theta^*, F^{obs} - \theta^* \rangle &= \langle \hat{\theta} - \tilde{\theta}, F^{obs} - \theta^* \rangle + \langle \tilde{\theta} - \theta^*, F^{obs} - \theta^* \rangle \\
&\leq \|\hat{\theta} - \tilde{\theta}\| \left| \left\langle \frac{\hat{\theta} - \tilde{\theta}}{\|\hat{\theta} - \tilde{\theta}\|}, F^{obs} - \theta^* \right\rangle \right| \qquad (4.4) \\
&\quad + \left(\|\tilde{\theta} - \hat{\theta}\| + \|\hat{\theta} - \theta^*\| \right) \left| \left\langle \frac{\tilde{\theta} - \theta^*}{\|\tilde{\theta} - \theta^*\|}, F^{obs} - \theta^* \right\rangle \right|. \qquad (4.5)
\end{aligned}$$

Using Lemmas XXX, the following three terms:

$$\|\hat{\theta} - \tilde{\theta}\|, \quad \left| \left\langle \frac{\hat{\theta} - \tilde{\theta}}{\|\hat{\theta} - \tilde{\theta}\|}, F^{obs} - \theta^* \right\rangle \right|, \quad \left| \left\langle \frac{\tilde{\theta} - \theta^*}{\|\tilde{\theta} - \theta^*\|}, F^{obs} - \theta^* \right\rangle \right| \tag{4.6}$$

can all be bounded by XXX with probability at least

$$XXX.$$

Combining these bounds with (4.4), (4.5) and (4.3), we get

$$\|\hat{\theta} - \theta^*\|^2 \leq XXX$$

with probability at least XXX.

Now we present the lemmas, which bound the three terms in (4.6), respectively.

Lemma 1. *For any constant $C' > 0$, there exists a constant $C > 0$ only depending on C', such that*

$$\|\hat{\theta} - \tilde{\theta}\| \leq CXXX,$$

with probability at least XXX.

Lemma 2. *For any constant $C' > 0$, there exists a constant $C > 0$ only depending on C', such that*

$$\left| \left\langle \frac{\tilde{\theta} - \theta^*}{\|\tilde{\theta} - \theta^*\|}, F^{obs} - \theta^* \right\rangle \right| \leq CXXX,$$

with probability at least XXX.

Lemma 3. *For any constant $C' > 0$, there exists a constant $C > 0$ only depending on C', such that*

$$\left| \left\langle \frac{\hat{\theta} - \tilde{\theta}}{\|\hat{\theta} - \tilde{\theta}\|}, F^{obs} - \theta^* \right\rangle \right| \leq CXXX,$$

with probability at least XXX.

□

5. Simulation

In the first case we analyze the network with two types of nodes. Figure 1 shows the locations of 50 nodes, among which 4 are from type I and the rest 46 are from type II. The first type of nodes (type I) are generated uniformly from $[0.3, 0.7] \times [0.8, 5.2]$. The second type of nodes (type II) are generated uniformly in $[0, 1] \times [0, 6]$.

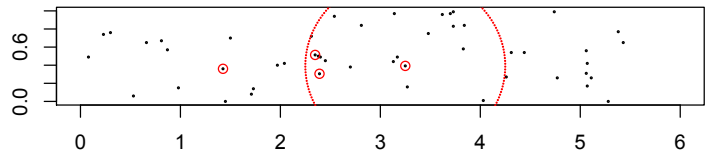

Fig 1: Case 1. Locations of nodes. Each black dot represents a node, the four type I nodes are marked as red. The dotted circle explains the reachable area of the second type I node.

The network is developed during time period $[0, 50]$. Two nodes are connected if the distance between them is less than 1. The connecting time between two type II nodes is generated from uniform distribution $U(0, 40)$. For the pair of nodes with one from type I and the other from type II, the connecting time is distributed as $N(5 + \tau, 1)$, where τ is the time delay caused by the type I node and is generated randomly from $U(0, 30)$.
Clustering results for one trial are shown in Fig 2.

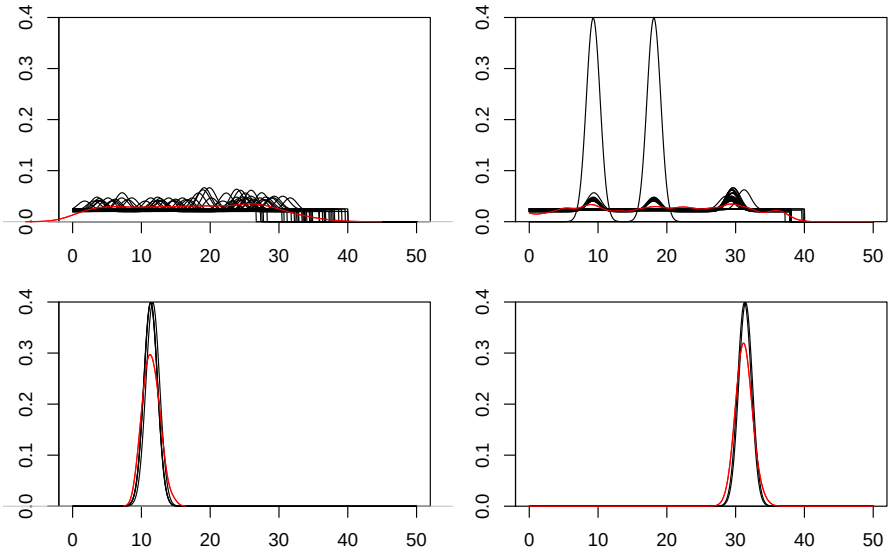

Fig 2: Estimated mean intensify functions for Case 1. Left column: estimates using c.d.f. Right column: estimates using smoothed p.d.f. Each column displays the results for two clusters. Red curves are the estimated mean intensity functions, black ones are the (aligned) true intensify functions which are assigned to that cluster. ["intensify function" is NOT rigorous. Change it later.]

The k-means++ initialization method is performed. For each trial, three independent initializations are conducted, and only the one leading to the largest between-cluster distance is kept, where the between-cluster distance is defined as the minimum pairwise distance between estimated mean functions. The clustering result is measured using adjusted rand index (ARI), which is also used in [6]. The algorithm is tested in 100 synthetic networks. The ARI is compared between trials based on c.d.f. and those based on smoothed p.d.f., and the result is shown in Figure 3.

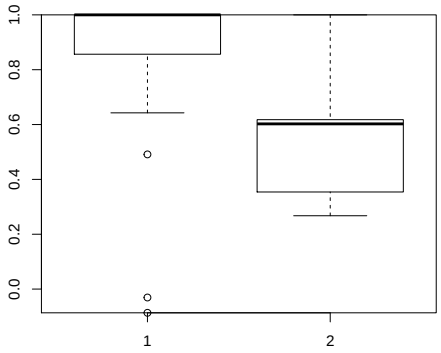

Fig 3: Case 1. ARI for trials using c.d.f (left) and p.d.f. (right).

The second case is with three clusters. The locations of nodes are displayed in Figure 4. The connecting radius for type I node is set as 2, and that for other nodes is set as 1. For a pair of type III nodes, the connecting time is generated from uniform distribution $U(0, 30)$. For the pair of nodes with one from type II and the other from type III, the connecting time is distributed as $N(5 + \tau, 1)$, where τ is the time delay caused by the type II node and is generated randomly from $U(0, 5)$. For the pair of nodes with one from type I and the other from type II, the connecting time is distributed as $U(\tau, \tau + 6)$, where τ is the time delay caused by the type II node and is generated randomly from $U(40, 42)$.

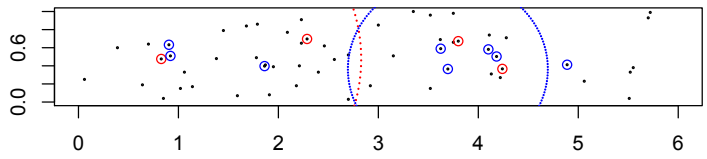

Fig 4: Case 2. Locations of nodes. Each black dot represents a node, the four type I nodes are marked as red, and eight type II nodes are marked as blue. The dotted red and blue circles explain the reachable area of the type I node and type II node, respectively.

Clustering results for one trial are shown in Fig 5.

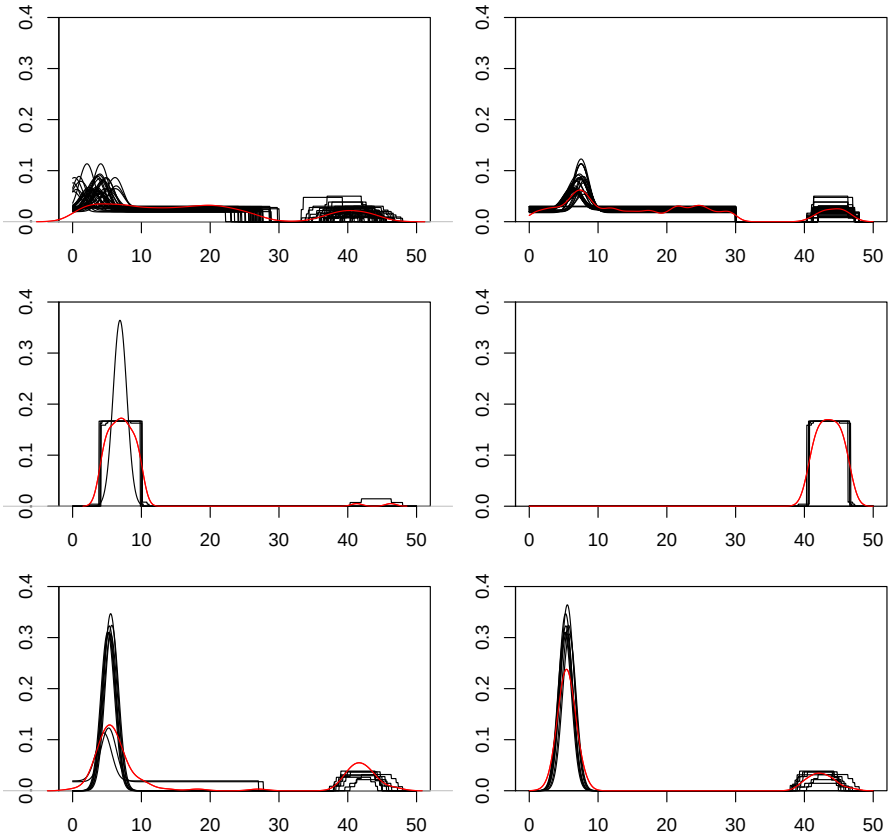

Fig 5: Estimated mean intensify functions for Case 2. Left column: estimates using c.d.f. Right column: estimates using smoothed p.d.f. Each column displays the results for three clusters. Red curves are the estimated mean intensity functions, black ones are the (aligned) true intensify functions of nodes which are assigned to that cluster. ["intensify function" is NOT rigorous. Change it later.]

100 independent trials are recorded. The k-means++ initialization is applied for each trial. The ARI is displayed in Figure 6.

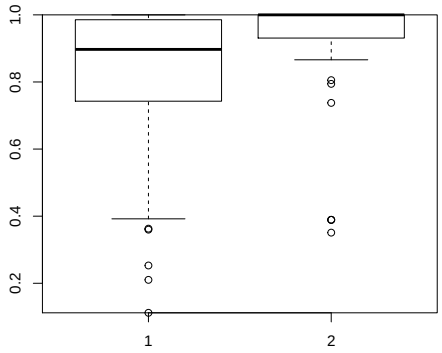

Fig 6: Case 2. ARI for trials using c.d.f (left) and p.d.f. (right).

The third case has similar set up as Case 2, except that the time delays of the type II nodes are generated from $U(0, 30)$ (it was $U(0, 5)$ in the second case). Clustering results for one trial are shown in Fig 7.

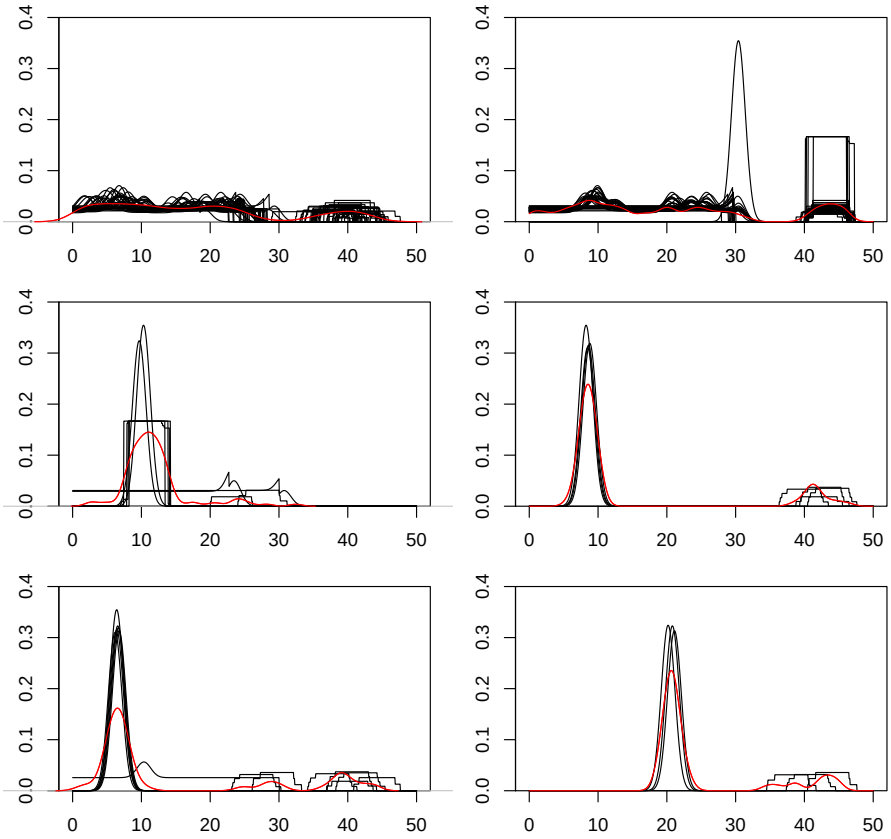

Fig 7: Estimated mean intensify functions for Case 3. Left column: estimates using c.d.f. Right column: estimates using smoothed p.d.f. Each column displays the results for three clusters. Red curves are the estimated mean intensity functions, black ones are the (aligned) true intensify functions of nodes which are assigned to that cluster. ["intensify function" is NOT rigorous. Change it later.]

100 independent trials are recorded. The k-means++ initialization is applied for each trial. The ARI is displayed in Figure 8.

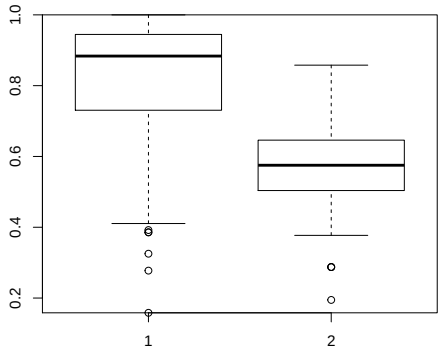

Fig 8: Case 3. ARI for trials using c.d.f (left) and p.d.f. (right).

[Will add comparison of different initialization strategies.]

6. Conclusion

Appendix A: Appendix section

TO DO:

1. Introduction.
2. Slides.
3. Visualization: compare initialization strategies, jitter boxplot (or violin)
4. Reading: basic knowledge about point process, and shift operation, identifiability.
5. Literature review.
6. Theory.
7. Real data: nodes location, etc.

A.1. Plots

Acknowledgements

See Supplement A for the supplementary material example.

Supplementary Material

Supplement A: Title of the Supplement A
(http://www.e-publications.org/ims/support/dowload/imsart-ims.zip). Dum esset rex in accubitu suo, nardus mea dedit odorem suavitatis. Quoniam confortavit seras portarum tuarum, benedixit filiis tuis in te. Qui posuit fines tuos

References

[1] Bachem, O., Lucic, M., Hamed Hassani, S. and Krause, A. (2017). Uniform Deviation Bounds for k-Means Clustering.

[2] Bigot, J. and Gendre, X. (2013). Minimax properties of Fréchet means of discretely sampled curves. *Ann. Stat.* **41** 923–956.

[3] Gao, C., Lu, Y. and Zhou, H. H. (2015). Rate-optimal graphon estimation. *Ann. Stat.* **43** 2624–2652.

[4] Holland, P. W., Blackmond, K. and Leinhardt, S. (1983). STOCHASTIC BLOCKMODELS: FIRST STEPS * Educational Testing Seroice ** Technical Report.

[5] Matias, C. and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **79** 1119–1141.

[6] Matias, C., Rebafka, T. and Villers, F. (2018). A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika* **105** 665–680.

[7] Pensky, M. (2019). Dynamic network models and graphon estimation. *Ann. Stat.* **47** 2378–2403.

[8] Pollard, D. (1981). Strong Consistency of K-Means Clustering. *Source Ann. Stat.* **9** 135–140.

[9] Wan, Y., Wei, Z., Looger, L. L., Koyama, M., Druckmann, S., Keller Correspondence, P. J. and Keller, P. J. (2019). Single-Cell Reconstruction of Emerging Population Activity in an Entire Developing Circuit. *Cell* **179**.

[10] Xu, K. S. (2015). Stochastic block transition models for dynamic networks. *J. Mach. Learn. Res.* **38** 1079–1087.

[11] Xu, K. S. and Hero, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE J. Sel. Top. Signal Process.* **8** 552–562.

[12] Yang, T., Chi, Y., Zhu, S., Gong, Y. and Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks - A Bayesian approach. *Mach. Learn.* **82** 157–189.