November 6, 2019

1 Introduction

2 Model

2.1 SBM

Denote the p nodes by $n_1, n_2, \cdots, n_p$. Let $Z_i \in \{1, 2, \cdots, K\}$ be the cluster that node n_i belongs to, where K is the number of clusters. Denote by $C_{k\times k}$ the connecting probability matrix, where $C_{kl} := P(n_i, n_j \text{ are connected} | Z_i = k, Z_j = l)$. The observed adjacency matrix $A_{p\times p}$ is defined as

$$A_{ij} = \begin{cases} 1, n_i \text{ and } n_j \text{ are connected;} \\ 0, \text{otherwise.} \end{cases}$$

SBM models this matrix by Bernoulli distribution, that is $A_{ij} \sim \text{Bernoulli}(C_{Z_i Z_j})$.

2.2 Dynamic Generalization of SBM

Under the assumption that the cluster is static and the edges do not disappear once constructed, the sequence of adjacency matrices $A(t)$ can be uniquely determined by $T_{p\times p}$ with entries

$$T_{ij} = \min \{t : A_{ij}(t) = 1\}.$$

On the other hand, one can model the observed adjacency matrices using point processes with intensity function

$$C_{kl}(t) = \frac{P\left(\mathrm{d}A_{ij}(t) = 1 | A_{ij}(t^-) = 0, Z_i = k, Z_j = l\right)}{\mathrm{d}t}$$

where $\mathrm{d}A_{ij}(t) = A_{ij}(t + \mathrm{d}t) - A_{ij}(t)$.

If we estimate the clustering vector Z based on the observed connecting time matrix T, local ensembles will be recognized. More specifically, the peripheral neurons that are connected to the same MN might be grouped together, and the neuron they are connecting with can then be clustered as a MN.

One could also estimate Z based on the intensity $\Lambda_i(t)$ of n_i. $\Lambda_i(t)$ has an explicit expression

$$\begin{aligned} \Lambda_i(t) &= P\left(\sum_{j=1}^{p} \mathrm{d}A_{ij}(t) \geq 1\right) / \mathrm{d}t \\ &= \sum_{j=1}^{p} P(\mathrm{d}A_{ij}(t) = 1)/\mathrm{d}t \\ &= \sum_{j=1}^{p} P(A_{ij}(t^-) = 0) \cdot C_{Z_i Z_j}(t) \\ &= \sum_{k=1}^{K} w_{Z_i k} \cdot C_{Z_i k}(t), \end{aligned}$$

where $w_{Z_i k} = \sum_{j:Z_j=k} P(A_{ij}(t^-) = 0)$.
Clustering based on $\Lambda_i(t)$ yields clusters corresponding to cell types.

3 Method

4 Theory

5 Simulation

5.1 Network with Two Cell Types

We analyze the network with two cell types. Thirty nodes $n_1, \cdots, n_{30}$ are generated uniformly at random in $[0, 1] \times [0, 1]$ so that the hazard function is constant. The first three nodes n_1, n_2, n_3 are labeled as "MNs" and others are labeled as "Others".
For each pair of nodes, the time of building a connection (an edge) is (independently) determined by their clusters. For the nodes in the same cluster, the connecting time is generated from the exponential distribution $Exp(0.1)$. For the nodes in different clusters, the connecting time points generated from both $Gamma(0.25, 0.1)$ and $Gamma(0.5, 0.1)$ are tested. The networks under these two settings will be referred to as $\mathbf{G}_1$ and $\mathbf{G}_2$ later on.
The time period of observation is set as $[0, 100]$, thus the connecting times are truncated at 100.

5.2 Network with Three Cell Types

We also ran simulation on a network with three cell types. Thirty nodes are located the same as above. The first three nodes n_1, n_2, n_3 are assigned to "Type 0", the following four $n_4, \cdots, n_7$ are assigned to "Type 1", and the rest are labeled as "Others".
Connecting time for nodes within the same cluster is distributed the same as above. Nodes of type "Type 0" build edges with other types of nodes at time points generated from $Gamma(0.25, 0.1)$. Nodes of "Type 1" build edges with "Other" nodes at time points generated from $N(1, 1)$.
This network will be referred to as $\mathbf{G}_3$.

5.3 Clustering Results

To show the development of the network, we plot snapshots in Fig.1 of the network $\mathbf{G}_1$ at time points $t = 0.01, 0.1, 1, 10, 100$.

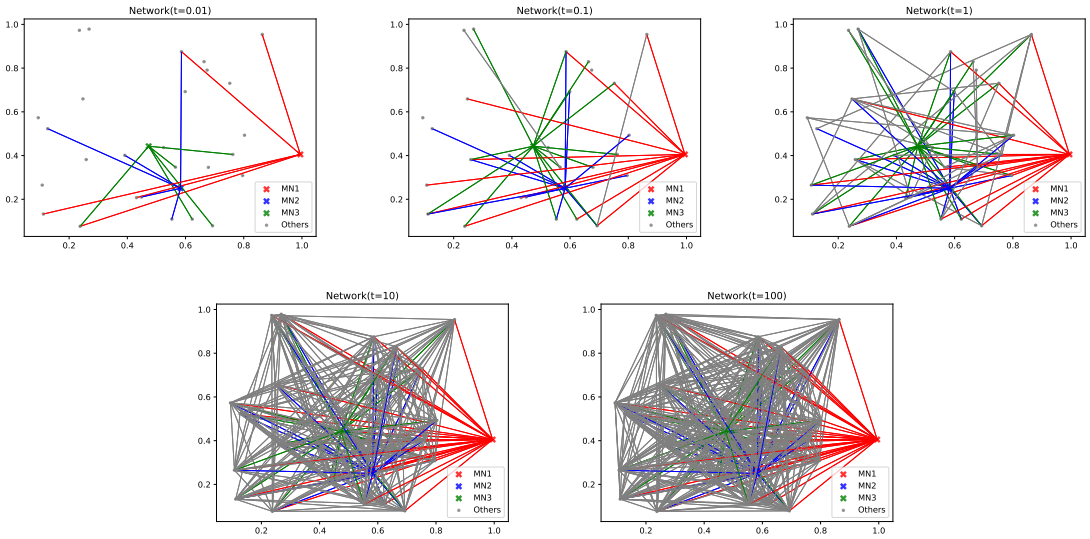

Figure 1: Development of the network with two cell types. The node n_1 is represented by "MN1" in red. The node n_2 is represented by "MN2" in blue. The node n_3 is represented by "MN3" in green. Other nodes are represented by "Others" in gray.

Kernel method is used to estimate the intensity function (???) of each node based on the connecting time points. Cross-validation method is applied in order to choose the optimal bandwidth.
The ℓ_2-distance is adopted as a measure of dissimilarity between the estimated intensity functions. The hierarchical clustering is then used to cluster nodes based on the pairwise ℓ_2-distances. Specifically, the average distance is used as the distance between two sets.
It turns out that $\mathbf{G}_1$ and $\mathbf{G}_3$ are clustered completely correct, whereas in $\mathbf{G}_2$ only n_3 is recognized successfully among the three "MNs".
To illustrate the reason, we plot the estimated intensity functions for each network (see Fig.2). It can be seen that the clusters of $\mathbf{G}_2$ are not distinct enough, which causes the failure to correctly cluster n_1 and n_2.

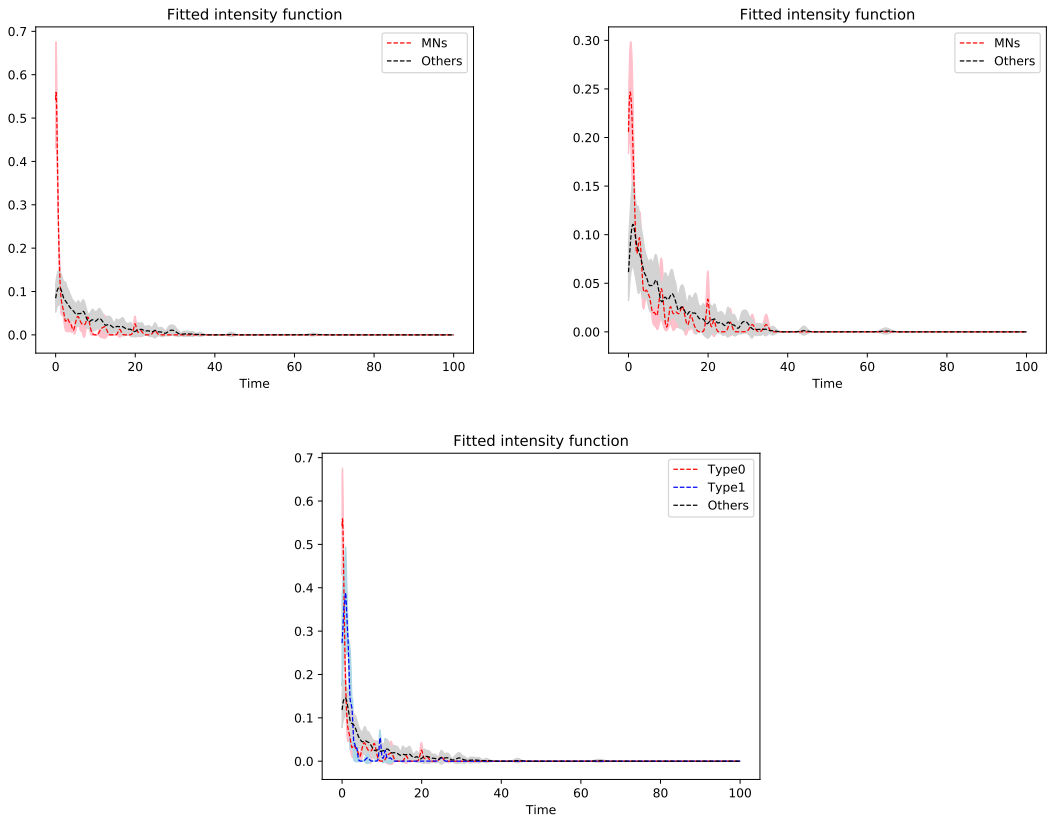

Figure 2: Fitted intensity functions. The top left figure corresponds to $\mathbf{G}_1$. The top right figure corresponds to $\mathbf{G}_2$. The bottom left figure corresponds to $\mathbf{G}_3$. The dashed lines represent the mean intensity of nodes in that cluster. The shaded area represents the standard deviations.

References