Random Shades of Colors: Multilayer clustering and community detection in networks

Brenda Betancourt,[1] Daniele Durante[2], and Rebecca C. Steorts[1]
Duke University[1]; University of Padua[2]

Summary

- We developed a model-based procedure that borrows information across layers to detect communities of nodes.
- Bayesian hierarchical model based on stochastic block modeling (SBM).
- Cluster membership of each node possibly vary across different layers.

Algorithm Description

The clustering mechanism can be described as a simple procedure based on random shades of colors (RASHAD):

- First stage: Assigns a specific color to each layer. Layers with the same color share the same community patterns.
- Second stage: Aggregates nodes and assigns a shade of color. Nodes with the same shade of color belong to same community.

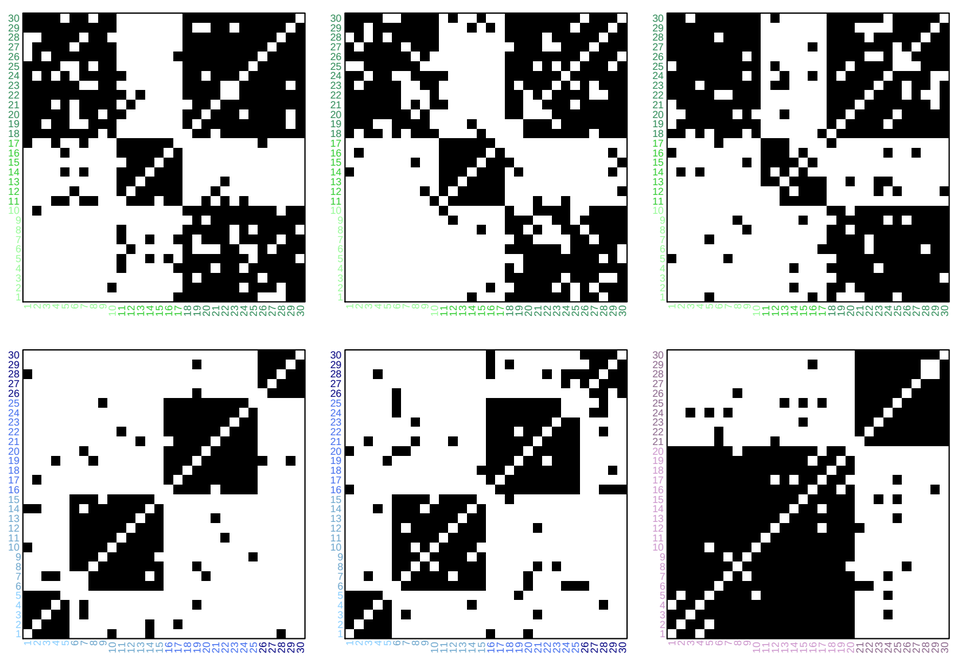

FIGURE 1: Adjacency matrices for simulated data set of 30 nodes with 6 layers of observations. The colors green, blue and purple represent the clustering of the layers.

RASHAD Model

Let $A_1, \ldots, A_N$ be the sequence of binary adjacency matrices representing the multilayer undirected network data set with no self-edges.

- Each layer $i = 1, \ldots, N$ consists of a $V \times V$ symmetric matrix A_i such that $A_{i[v,u]} = A_{i[u,v]} = 1$ if there is a connection between nodes v and u in layer i, and 0 otherwise.
- The cluster assignments $c_1, \ldots, c_N$ with $c_i \in \{1, \ldots, K\}$ indicate the color of layer i.
- The color-specific partition of the node set is obtained via $s_{1k}, \ldots, s_{Vk}$, with $s_{vk} \in \{1, \ldots, H_k\}$ the shade of node v in color k and H_k the total number of communities in color k.

This construction leads to the following generative mechanism:

$$A_{i[v,u]} = A_{i[u,v]} \mid c_i = k \sim \text{Bern}(\pi_{k[v,u]}), \quad (1)$$
$$\pi_{k[v,u]} \mid s_{vk} = h, s_{uk} = h' = \theta_{k[h,h']}. \quad (2)$$

Prior Specification

- The prior on interaction probabilities is: $\theta_{k[h,h']} \overset{\text{iid}}{\sim} \text{Beta}(a, b)$
- Chinese Restaurant Processes (CRP) for the cluster assignments at both levels allows the number of clusters to be random:

$$c = (c_1, \ldots, c_N) \sim \text{CRP}(\alpha_c),$$
$$s_k = (s_{1k}, \ldots, s_{Vk}) \sim \text{CRP}(\alpha_k).$$

The prior distribution over colors for the ith layer, conditioned on the colors of the others $c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_N$ is

$$\text{pr}(c_i = k \mid \ldots) = \begin{cases} \frac{n_{k,-i}}{N-1+\alpha_c} & \text{for } k = 1, \ldots, K_{-i}, \\ \frac{\alpha_c}{N-1+\alpha_c} & \text{for } k = K_{-i}+1, \end{cases}$$

where $n_{k,-i}$ is the total number of layers associated to color k.

Similarly, for the shading mechanism:

$$\text{pr}(s_{vk} = h \mid \ldots) = \begin{cases} \frac{m_{hk,-v}}{V-1+\alpha_k} & \text{for } h = 1, \ldots, H_{k,-v}, \\ \frac{\alpha_k}{V-1+\alpha_k} & \text{for } h = H_{k,-v}+1, \end{cases}$$

independently, for each $k = 1, \ldots, K$, where $m_{hk,-v}$ denotes the total number of nodes with shade h of color k, when node v is held out.

- Gamma hyperpriors on α_c and α_k.

Posterior samples are obtained via collapsed Gibbs sampler marginalizing over the interaction probabilities and results of Escobar and West (1995).

Application

- We use real data from an Indian village in the state of Karnataka.
- Six types of relationships (layers) between the 114 households (nodes) of the village.
- Sparse networks with an average percent of links of 2.4% except for the temple relationship with only 0.25%.
- The model identifies attending the same temple as the only relationship with different community structure across households.

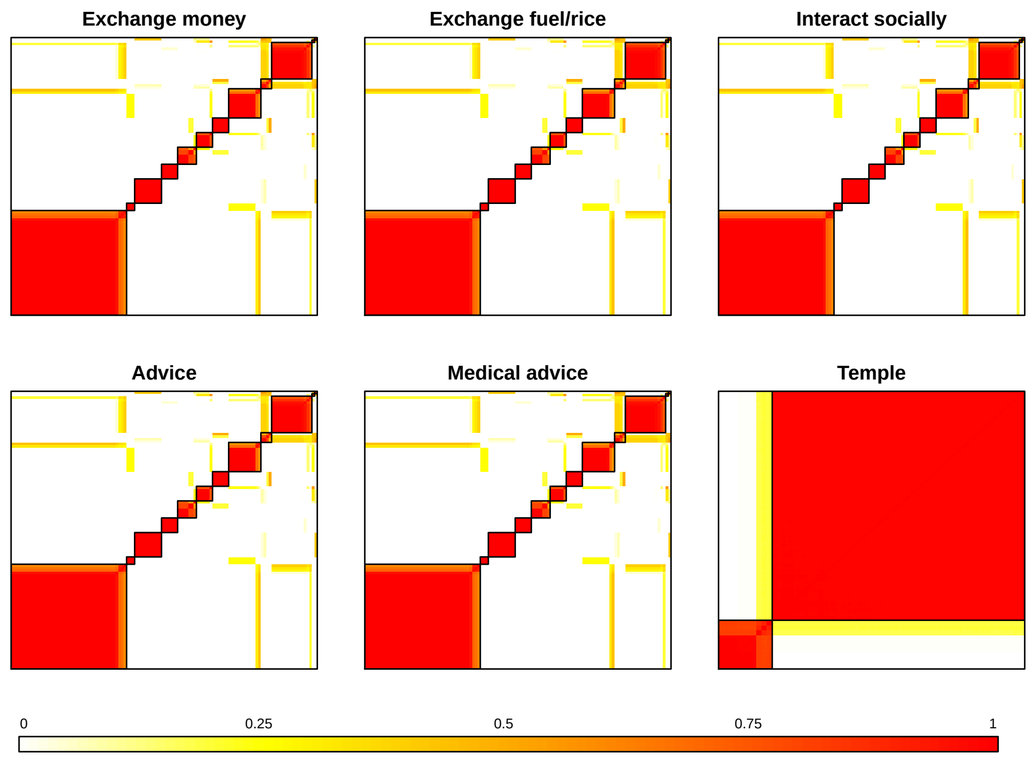

FIGURE 2: Mean posterior pairwise incidence matrices for layers of Karnataka village. Red represents a high probability of two households to have the same shade of color within each layer.

Discussion

- We developed an efficient algorithm which allows automatic learning of the total number of clusters at the layer and nodes levels.
- RASHAD can easily accommodate other type of data and priors for random partitions such as Pitman-Yor process.

Acknowledgements: Participation in this conference was supported in part by the WiML travel grant. This work is supported by the Foester-Bernstein Postdoctoral Fellowship.