Stochastic block models for community detection in heterogeneous networks

Hamed Kabiri Kenari

Kharazmi University, Iran

Abstract

Heterogeneous networks have multiple types of nodes and edges. Community detection in single layer and multiplex networks has been extensively studied in the past decade. But there are few methods have constructed for heterogeneous networks. In this paper, we introduce heterogeneous stochastic block models for detecting communities in heterogeneous networks. Generally these models are developed based on generalization of single-layer stochastic block model, bipartite stochastic block model and multiplex stochastic block model. We define this two types of stochastic block model, independent degree and shared degree. Independent degree models have one specific degree parameter for each layers, shared degree models share one degree parameter for all layers. We introduce a method to create synthetic networks with benchmark heterogeneous communities. We evaluate the performance of the proposed community detection algorithm with generalization of Kernighan-Lin algorithm in the controlled environment (with synthetic benchmark communities). According to our results, shared degree models have better performance in high crossed networks in contrast independent degree models have better performance in low crossed networks. Exception when intra-layer densities are high and inter-layer densities are low, single-layer algorithm (flattering network) has better performance. On real datasets, DBLP and AMiner four-area datasets, proposed methods have good results.

Keywords: Heterogeneous Networks, Stochastic Block Models, Community Detection

1. Introduction

Network community detection has very useful application in different scientific fields, such as physics, biology, statistics, information technology, social science and many others. Many real networks are heterogeneous that have different types of nodes and edges.

Definition of is that heterogeneous networks have multiple types of nodes and edges. For example, in a healthcare network, nodes can be patients, diseases, doctors and hospitals. The edges can be in the type of patient-disease (patient treated for disease), patient-doctor (patient treated by doctor), doctor-hospital (doctor works at hospital). Figure 1 shows a heterogeneous network. It shows location based social network (say, Yelp) as heterogeneous network, there are two types of nodes, users and locations. Furthermore, there are three types of interactions in this network. A user is linked to another user through friendship and a location node is connected through proximity links to other locations. Also location node represents the visit of a customer to a location.

We can interpret heterogeneous networks as multilayer networks that have some layers each one has one type of nodes and edges connect them and some bipartite networks each one interconnects different types of nodes. Detecting communities algorithms in heterogeneous networks have been introduced recently in (Sun and Han, 2021; X. Liu and Wakita, 2014; Zhang and Chen, 2018; Sengupta and Chen, 2015; Li et al., 2016). Simple approach to community detection in heterogeneous networks is applying standard community detection algorithms to a one layer (flattering) network. But flattering approach causes we lose many useful information about heterogeneous networks and community detection algorithms can not find best results. For solving this issue, heterogeneous version of Girvan-Newman modularity have been proposed in (D. Liu and Ma, 2020; Song et al., 2015). But there is no extension of stochastic block models (SBMs). We want to develop heterogeneous version of SBM.

Email address: hamedforphd@email.com (Hamed Kabiri Kenari)

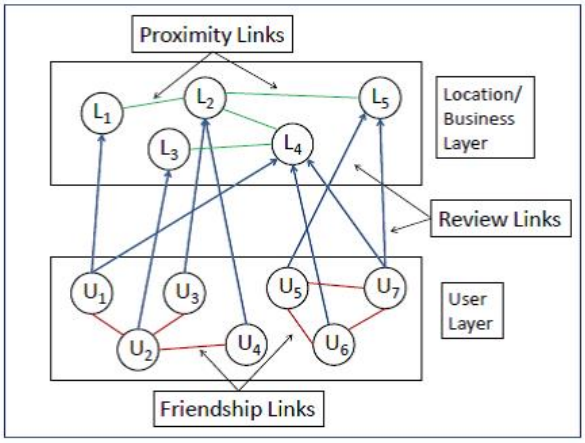

Figure 1: A sample multilayer (Yelp) network. (Pramanik et al., 2017)

Stochastic block models (SBMs) are graceful probabilistic models of community structure in networks that have some different kind. Degree corrected SBM in (Karrer and Newman, 2011) introduced and in (Larremore et al., 2014) introduced bipartite SBM and in (Valles-Catala et al., 2016; Han et al., 2015; Peixoto, 2015; Paul and Chen, 2015; Taylor et al., 2016; Stanley et al., 2015) introduced multilayer SBMs.

In (Paul and Chen, 2016) developed several version of multilayer stochastic models. There models are for multilayer homogenous networks. We can categorize them to two: independent degree, shared degree. Also they used restricted MLSBMs to develop more models. We use their categories and introduce new models for heterogeneous network. Independent degree and shared degree and restricted version of them are proposed. For restricted version we define three different versions because we have two different intra layers and inter layers that we can define restricted version on them.

2. Definitions

First, we introduce a definition of heterogeneous multilayer networks because the new definition is more useful for generating data in the future.

We define a heterogeneous network as $G = (G_U, G_B)$ where $G_U = \{L_i : i \in \{1, 2, \ldots, M\}\}$ is a set of M graphs and $G_B = \{L_{ij} : i, j \in \{1, 2, \ldots, M\}, i \neq j\}$ is a set of bipartite graphs that contain of intra-layers and inter-layer edges. Each layer $L_i = (V_i, E_i)$, with $V_i$and $E_i$represent the nodes and inner edges of each layer, respectively. Similarly, we can have $L_{ij} = (V_i, V_i, E_{ij})$where $\{E_{ij} \subseteq V_i \times V_j : i, j \in \{1, 2, \ldots, M\}, i \neq j\}$ a bipartite graph that pairs between layers $L_i$and L_j. Here we define community in heterogeneous multilayer networks.

The community C in a heterogeneous multilayer network G is defined as a crossed module $(C_U, C_B)$of G containing a subset of single or multilayer nodes and all edges between them. Mathematically, C_U and $C_B$can be expressed as $C_U = \left\{L_i^C = \left(V_i^C, E_i^C\right) : V_i^C \subseteq V_i, E_i^C = \left\{E_i \cap \left(V_i^C, V_i^C\right)\right\}, i \in \{1, 2, \ldots, M\}\right\}$ and $C_B = \left\{L_{ij}^C = \left(V_i^C, V_j^C, E_{ij}^C\right) : V_i^C \subseteq V_i, V_j^C \subseteq V_j, E_{ij}^C = \left\{E_{ij} \cap \left(V_i^C \times V_j^C\right)\right\}, i, j \in \{1, 2, \ldots, M\}, i \neq j\right\}$. Importantly, the communities of a multilayer network G can be divided into two types (see Figure 2). (a) cross-layer communities (containing several types of nodes) that $|C_B| \neq K$ (b) single-layer communities (containing only a single type of nodes) that $|C_U| = K$. In the following sections we formulate the heterogeneous stochastic block models and describe an algorithm that searches for a maximum likelihood partition of a network into communities. We first compare these models can on synthetic network partition. We then apply these models to several empirical networks, showing that these models outperform their one layer SBM counterpart.

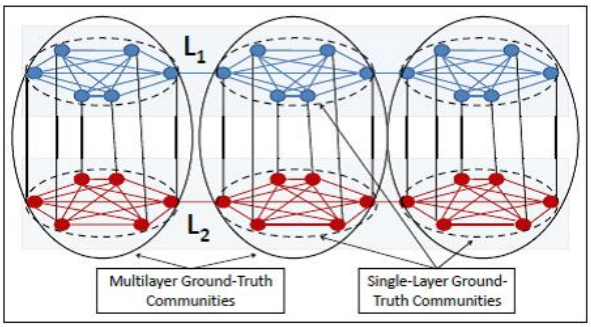

Figure 2: Network configurations with different bilayers. (Pramanik et al., 2017)

3. Heterogeneous Multilayer Stochastic Block Model

Throughout the section, we assume that networks have N nodes and M different types, adjacency matrix A is $N \times N$ and the edges $A_{ij}^{ml}$are located between two nodes i of layer m and node j of layer l, formed independently the Poisson distribution. Similarly, we express the matrix of group interrelationships $\pi^{m,l}$ as a $K \times K$ matrix .The communities associated with nodes are $z_i$and $z_j$and the degree vectors of nodes are $k_i^{ml}$and $k_j^{lm}$and let $\theta_i^{m,l}$ control the expected degree of vertex i:

$$A_{ij}^{ml} | \left(z_i = q, z_j = s\right) = A_{ji}^{lm} | \left(z_j = s, z_i = q\right) \sim Poisson(P_{ij}^{m,l})$$

We will use the notations $e_{qs}^{m,l}$ to denote the total number of edges between communities q in layer m and community s in layer l. We will use the notations $e_q^{m,l}$ to denote the total number of edges in communities q between layer m and layer l , i.e., $e_{qs}^{m,l} = \sum_{i,j} A_{ij}^{m,l} I\left(z_i = q, z_j = s\right)$ and $e_q^{m,l} = \sum_{i,j} A_{i,j}^{m,l} I\left(z_i = q\right) = \sum_i k_i^{m,l} I\left(z_i = q\right)$, where $I\left(.\right)$ Is the indicator function which is 1 if the condition inside is satisfied and 0 otherwise. Note that $e_{qs}^{m,l} = e_{sq}^{m,l}$, $e_q^{m,l} = \sum_s e_{qs}^{m,l}$ and $e_{qq}^{m,l}$ is twice the number of edges within the community q between layer m and l .

We model the Poisson mean parameter for multilayer random block model in five different ways with number of different variables.

The first model is an independent degree model. We define:

$$P_{ij}^{m,l} = \theta_i^{m,l} \theta_j^{m,l} \pi_{qs}^{m,l} \quad , \quad i,j \in \{1,\dots N\}, \quad m,l \in \{1,\dots M\} \quad q,s \in \{1,\dots,K\}$$

With restrictions:

$$\sum_{i:z_{i=q}} \theta_i^{m,l} = 1, \quad m,l \in \{1,\dots,M\}, \quad q \in \{1,\dots,K\}$$

We call it degree corrected multilayer stochastic Block Model (DCMLSBM).

The next model is the modified version of the RMLSBM, which we call it DCRMLSBM-intra-layer. In this model we have for intra layers:

$$P_{ij}^{m,l} = \theta_i^{m,l} \theta_j^{m,l} \pi_{qs} \quad , \quad i,j \in \{1,\dots N\}, \quad m,l \in \{1,\dots M\}, \quad m = l \quad q,s \in \{1,\dots,K\}$$

With restrictions:

$$\sum_{i:z_{i=q}} \theta_i^{m,l} = 1, \quad m,l \in \{1,\dots,M\}, \quad q \in \{1,\dots,K\}$$

The next model is the modified version of the RMLSBM degree, which we call DCRMLSBM-inter-layer. In this model we have for inter layers:

$$P_{ij}^{m,l} = \theta_i^{m,l} \theta_j^{m,l} \pi_{qs} \quad , \quad i,j \in \{1,\dots N\}, \quad m,l \in \{1,\dots M\}, \quad m \neq l \quad q,s \in \{1,\dots,K\}$$

With restrictions:

$$\sum_{i:z_{i=q}} \theta_i^{m,l} = 1, \quad m,l \in \{1,\dots,M\}, \quad q \in \{1,\dots,K\}$$

The next model is the modified version of the RMLSBM degree, which we call DCRMLSBM. In this model we have for all layers:

$$P_{ij}^{m,l} = \theta_i^{m,l}\theta_j^{m,l}\pi_{qs} \quad , \quad i,j \in \{1,\dots N\}, \ m,l \in \{1,\dots M\}, \qquad q,s \in \{1,\dots,K\}$$

With restrictions:

$$\sum_{i:z_{i=q}} \theta_i^{m,l} = 1, \quad m,l \in \{1,\dots,M\}, \quad q \in \{1,\dots,K\}$$

In the next model, the underlying model is shared degree, and hence the specific degree parameter of each node is shared across the layers. We call this model the Block Shared Degree Restricted Multi-Layer Stochastic Block Model (SDRMLSBM). The model can be written as:

$$P_{ij}^{m,l} = \theta_i^m\theta_j^l\beta_{m,l}\pi_{qs} \quad , \quad i,j \in \{1,\dots N\}, \ m,l \in \{1,\dots M\}, \qquad q,s \in \{1,\dots,K\}$$

With restrictions:

$$\sum_{m}\sum_{i:z_i=q} \theta_i^m = 1 \ \ m \in \{1,\dots,M\}, \quad q \in \{1,\dots,K\}$$

$$\sum_{m,l} \beta_{m,l} = 1$$

We use the likelihood method, which is similarly used in 8 and 9, so We maximize log-likelihood, $l(A|z;P)$, that obtained from the given communities and adjacent matrix. This is done by substituting in maximize likelihood estimation from the conditional parameter set P on z. log-likelihood condition of DCMLSBM can be written as (excluding sentences that have no role in the appointment of communities):

$$log-likelihood\,(A;z,\pi,\theta) = \sum_{m,l=1}^{M}\sum_{i<j}\left\{A_{ij}^{m,l}\left\{\log\left(\pi_{z_iz_j}^{m,l}\right) + \log\left(\theta_i^{m,l}\right) + \log(\theta_j^{m,l})\right\} - \theta_i^{m,l}\theta_j^{m,l}\pi_{z_iz_j}^{m,l}\right\}$$

$$= \sum_{m,l}\sum_{i} k_i^{m,l}\log\left(\theta_i^{m,l}\right) + \sum_{m,l}\sum_{q\le s}\left\{e_{qs}^{m,l}\log\left(\pi_{qs}^{m,l}\right) - \pi_{qs}^{m,l}\right\}$$

Maximize likelihood estimation for π can be obtained directly from the log-likelihood function. But for obtaining maximize likelihood estimation for θ under constraints we have to use Lagrange multiplications. As a result, we optimize the following objective function:

$$(\theta,\mu) = \sum_{i}\sum_{m,l} k_i^{m,l}\log\left(\theta_i^{m,l}\right) + \sum_{m,l}\sum_{q}\mu_q^{m,l}\Big(\sum_{i:z_i=q}\theta_i^{m,l} - 1\Big)$$

Solving maximize likelihood estimation gives the following values for θ and μ:

$$\widehat{\theta}_i^{m,l} = \frac{k_i^{m,l}}{\sum_{i:z_i=q} k_i^{m,l}} = \frac{k_i^{m,l}}{e_q^{m,l}}$$

$$\widehat{\pi}_{qs}^{m,l} = \sum_{i,j:z_i=q,z_j=s} A_{ij}^{m,l} = e_{qs}^{m,l}$$

Substituting these estimates into the log-likelihood function.

$$log-likelihood\,(A;z) = \sum_{i}\sum_{m,l} k_i^{m,l}\log\left(\frac{k_i^{m,l}}{e_q^{m,l}}\right) + \sum_{m,l}\sum_{q\le s}\left\{e_{qs}^{m,l}\log\left(e_{qs}^{m,l}\right) - e_{qs}^{m,l}\right\}$$

$$= \sum_{m,l} \sum_{q \leq s} e_{qs}^{m,l} \log\left(e_{qs}^{m,l}\right) - \sum_{m,l} \sum_{q \leq s} e_{qs}^{m,l}$$

$$+ \sum_{i} \sum_{m,l} k_i^{m,l} \log\left(k_i^{m,l}\right) - \sum_{q} \sum_{m,l} e_q^{m,l} log\left(e_q^{m,l}\right)$$

Now we have ignoring the sentences that do not depend on the appointment of communities (sentences 2 and 3)

$$log - likelihood\,(A; z) = \sum_{m,l} \sum_{q \leq s} e_{qs}^{m,l} \log\left(e_{qs}^{m,l}\right) - \sum_{q} \sum_{m,l} e_q^{m,l} log\left(e_q^{m,l}\right)$$

It is easy to understand that this maximum likelihood function can be written as follows

$$Q_{\mathrm{DCMLSBM}} = \sum_{m,l} \sum_{q \leq s} \left\{ e_{qs}^{m,l} \log\left(\frac{e_{qs}^{m,l}}{e_q^{m,l} e_s^{m,l}}\right) \right\}$$

Its normalized version based on each layers

$$Q_{\mathrm{DCMLSBM}} = \sum_{m,l} \sum_{q \leq s} \left\{ (e_{qs}^{m,l}/N^{m,l}) \log\left(\frac{(e_{qs}^{m,l}/N^{m,l})}{(e_q^{m,l}/N^{m,l})(e_s^{m,l}/N^{m,l})}\right) \right\}$$

Similarly, for DCRMLSBM-Intra-layer the conditional probability with constraints can be simplified as (omitting statements that are not parameter dependent).

$$log{-}likelihood\,(A; z, \pi, \theta) = \sum_{\substack{m,l=1 \\ m \neq l}}^{M} \sum_{i<j} \left\{ A_{ij}^{m,l} \left\{ \log\left(\pi_{z_i z_j}^{m,l}\right) + \log\left(\theta_i^{m,l}\right) + \log\left(\theta_j^{m,l}\right) \right\} - \theta_i^{m,l} \theta_j^{m,l} \pi_{z_i z_j}^{m,l} \right\}$$

$$+ \sum_{m=1}^{M} \sum_{i<j} \left\{ A_{ij}^{m,m} \left\{ \log\left(\pi_{z_i z_j}\right) + \log\left(\theta_i^{m,m}\right) + \log\left(\theta_j^{m,m}\right) \right\} - \theta_i^{m,m} \theta_j^{m,m} \pi_{z_i z_j} \right\}$$

$$= \sum_{\substack{m,l \\ m \neq l}} \sum_{i} k_i^{m,l} \log\left(\theta_i^{m,l}\right) + \sum_{\substack{m,l \\ m \neq l}} \sum_{q \leq s} \left\{ e_{qs}^{m,l} \log\left(\pi_{qs}^{m,l}\right) - \pi_{qs}^{m,l} \right\}$$

$$+ \sum_{m} \sum_{i} k_i^{m,m} \log\left(\theta_i^{m,m}\right) + \sum_{m} \sum_{q \leq s} \left\{ e_{qs}^{m,m} \log\left(\pi_{qs}\right) - \pi_{qs} \right\}$$

The θ and π in maximize likelihood estimation below and constraints are again obtained using the Lagrange method as previously described:

$$\widehat{\theta}_i^{m,l} = \frac{k_i^{m,l}}{\sum_{i:z_i=q} k_i^{m,l}} = \frac{k_i^{m,l}}{e_q^{m,l}}$$

$$\widehat{\pi}_{qs}^{m,l} = \sum_{i,j:z_i=q,z_j=s} A_{ij}^{m,l} = e_{qs}^{m,l}, \qquad m \neq l$$

$$\pi_{qs} = \sum_{m} \sum_{i,j:z_i=q,z_j=s} A_{ij}^{m,m} = \sum_{m} e_{qs}^{m,m}$$

The Modularity function can be obtained by pasting in maximize likelihood estimation and then deleting sentences that do not depend on community assignment:

$$log - likelihood\,(A; z) = \sum_{\substack{m,l \\ m \neq l}} \sum_{q \leq s} e_{qs}^{m,l} \log\left(e_{qs}^{m,l}\right) - \sum_{q} \sum_{\substack{m,l \\ m \neq l}} e_q^{m,l} log\left(e_q^{m,l}\right)$$

$$+\sum_{m}\sum_{q\leq s} e_{qs}^{m,m}\log\left(\sum_{m} e_{qs}^{m,m}\right) - \sum_{q}\sum_{m} e_{q}^{m,m} log\left(e_{q}^{m,l}\right)$$

It is easy to understand that this maximum likelihood function can be written as

$$Q_{\mathrm{DCRMLSBM-Intra-layer}} = \sum_{\substack{m,l \\ m\neq l}} \sum_{q\leq s}\left\{e_{qs}^{m,l}\log\left(\frac{e_{qs}^{m,l}}{e_{q}^{m,l}e_{s}^{m,l}}\right)\right\} + \sum_{m}\sum_{q\leq s}\left\{e_{qs}^{m,m}\log\left(\frac{\sum_{m} e_{qs}^{m,m}}{e_{q}^{m,l}e_{s}^{m,l}}\right)\right\}$$

Its normalized version based on each layers

$$Q_{\mathrm{DCRMLSBM-Intra-layer}} = \sum_{\substack{m,l \\ m\neq l}} \sum_{q\leq s}\left\{(e_{qs}^{m,l}/N^{m,l})\log\left(\frac{(e_{qs}^{m,l}/N^{m,l})}{(e_{q}^{m,l}/N^{m,l})(e_{s}^{m,l}/N^{m,l})}\right)\right\}$$

$$+\sum_{m}\sum_{q\leq s}\left\{(e_{qs}^{m,m}/N^{m,m})\log\left(\frac{\sum_{m}(e_{qs}^{m,m}/N^{m,m})}{(e_{q}^{m,m}/N^{m,m})(e_{s}^{m,m}/N^{m,m})}\right)\right\}$$

Similarly, for DCRMLSBM-Inter-layer the conditional probability with constraints can be simplified as (omitting statements that are not parameter dependent)

$$log-likelihood\left(A;z,\pi,\theta\right) = \sum_{\substack{m,l=1 \\ m\neq l}}^{M} \sum_{i<j}\left\{A_{ij}^{m,l}\left\{\log\left(\pi_{z_i z_j}\right) + \log\left(\theta_{i}^{m,l}\right) + \log(\theta_{j}^{m,l})\right\} - \theta_{i}^{m,l}\theta_{j}^{m,l}\pi_{z_i z_j}\right\}$$

$$+\sum_{m=1}^{M}\sum_{i<j}\left\{A_{ij}^{m,m}\left\{\log\left(\pi_{z_i z_j}^{m,m}\right) + \log\left(\theta_{i}^{m,m}\right) + \log(\theta_{j}^{m,m})\right\} - \theta_{i}^{m,m}\theta_{j}^{m,m}\pi_{z_i z_j}^{m,m}\right\}$$

$$= \sum_{\substack{m,l \\ m\neq l}}\sum_{i} k_{i}^{m,l}\log\left(\theta_{i}^{m,l}\right) + \sum_{\substack{m,l \\ m\neq l}}\sum_{q\leq s}\left\{e_{qs}^{m,l}\log\left(\pi_{qs}\right) - \pi_{qs}\right\}$$

$$+\sum_{m}\sum_{i} k_{i}^{m,m}\log\left(\theta_{i}^{m,m}\right) + \sum_{m}\sum_{q\leq s}\left\{e_{qs}^{m,m}\log\left(\pi_{qs}^{m,m}\right) - \pi_{qs}^{m,m}\right\}$$

The θ and π in maximize likelihood estimation below and the constraints are again obtained using the Lagrange method as previously described:

$$\widehat{\theta}_{i}^{m,l} = \frac{k_{i}^{m,l}}{\sum_{i:z_i=q} k_{i}^{m,l}} = \frac{k_{i}^{m,l}}{e_{q}^{m,l}}$$

$$\widehat{\pi}_{qs} = \sum_{m,l}\sum_{i,j:z_i=q,z_j=s} A_{ij}^{m,l} = \sum_{\substack{m,l \\ m\neq l}} e_{qs}^{m,l}, \qquad m\neq l$$

$$\pi_{qs}^{m,m} = \sum_{i,j:z_i=q,z_j=s} A_{ij}^{m,m} = e_{qs}^{m,m}$$

The Modularity function can be obtained by pasting in maximize likelihood estimation and then deleting sentences that do not depend on community assignment:

$$log-likelihood\,(A;z) = \sum_{\substack{m,l \\ m\neq l}} \sum_{q\leq s} e_{qs}^{m,l} \log\left(\sum_{\substack{m,l \\ m\neq l}} e_{qs}^{m,l} \right) - \sum_{q} \sum_{\substack{m,l \\ m\neq l}} e_{q}^{m,l} log\left(e_{q}^{m,l}\right)$$

$$+ \sum_{m} \sum_{q\leq s} e_{qs}^{m,m} \log\left(e_{qs}^{m,m}\right) - \sum_{q} \sum_{m} e_{q}^{m,m} log\left(e_{q}^{m,l}\right)$$

It is easy to understand that this maximum likelihood function can be written as

$$Q_{\mathrm{DCRMLSBM-Inter-layer}} = \sum_{\substack{m,l \\ m\neq l}} \sum_{q\leq s} \left\{ e_{qs}^{m,l} \log\left(\frac{\sum_{\substack{m,l \\ m\neq l}} e_{qs}^{m,l}}{e_{q}^{m,l} e_{s}^{m,l}} \right) \right\} + \sum_{m} \sum_{q\leq s} \left\{ e_{qs}^{m,m} \log\left(\frac{e_{qs}^{m,m}}{e_{q}^{m,l} e_{s}^{m,l}} \right) \right\}$$

Its normalized version based on each layers

$$Q_{\mathrm{DCRMLSBM-Inter-layer}} = \sum_{\substack{m,l \\ m\neq l}} \sum_{q\leq s} \left\{ (e_{qs}^{m,l}/N^{m,l}) \log\left(\frac{\sum_{\substack{m,l \\ m\neq l}} (e_{qs}^{m,l}/N^{m,l})}{(e_{q}^{m,l}/N^{m,l})(e_{s}^{m,l}/N^{m,l})} \right) \right\}$$

$$+ \sum_{m} \sum_{q\leq s} \left\{ (e_{qs}^{m,m}/N^{m,m}) \log\left(\frac{(e_{qs}^{m,m}/N^{m,m})}{(e_{q}^{m,m}/N^{m,m})(e_{s}^{m,m}/N^{m,m})} \right) \right\}$$

Similarly for DCRMLSBM the conditional probability with constraints can be simplified as (omitting statements that are not parameter dependent)

$$log-likelihood\,(A;z,\pi,\theta) = \sum_{m,l=1}^{M} \sum_{i<j} \left\{ A_{ij}^{m,l} \left\{ \log\left(\pi_{z_i z_j}\right) + \log\left(\theta_{i}^{m,l}\right) + \log(\theta_{j}^{m,l}) \right\} - \theta_{i}^{m,l} \theta_{j}^{m,l} \pi_{z_i z_j} \right\}$$

The θ and π in maximize likelihood estimation below the constraints are again obtained using the Lagrange method as previously described:

$$\widehat{\theta}_{i}^{m,l} = \frac{k_{i}^{m,l}}{\sum_{i:z_i=q} k_{i}^{m,l}} = \frac{k_{i}^{m,l}}{e_{q}^{m,l}}$$

$$\widehat{\pi}_{qs} = \sum_{m,l} \sum_{i,j:z_i=q,z_j=s} A_{ij}^{m,l} = \sum_{m,l} e_{qs}^{m,l},$$

The Modularity function can be obtained by pasting in maximize likelihood estimation and then deleting sentences that do not depend on community assignment:

$$log-likelihood\,(A;z) = \sum_{m,l} \sum_{q\leq s} e_{qs}^{m,l} \log\left(\sum_{m,l} e_{qs}^{m,l} \right) - \sum_{q} \sum_{m,l} e_{q}^{m,l} log\left(e_{q}^{m,l}\right)$$

It is easy to understand that this maximum likelihood function can be written as follows

$$Q_{\mathrm{DCRMLSBM}} = \sum_{m,l} \sum_{q \leq s} \left\{ e_{qs}^{m,l} \log \left(\frac{\sum_{m,l} e_{qs}^{m,l}}{e_q^{m,l} e_s^{m,l}} \right) \right\}$$

Its normalized version based on each layers

$$Q_{\mathrm{DCRMLSBM}} = \sum_{m,l} \sum_{q \leq s} \left\{ (e_{qs}^{m,l}/N^{m,l}) \log \left(\frac{\sum_{m,l} (e_{qs}^{m,l}/N^{m,l})}{(e_q^{m,l}/N^{m,l})(e_s^{m,l}/N^{m,l})} \right) \right\}$$

For SDRMLSBM, the log-likelihood function becomes conditional without parameter-independent statements

$$log - likelihood\,(A; z, \pi, \theta) = \sum_{m,l=1}^{M} \sum_{i<j} A_{ij}^{m,l} \left\{ \log \left(\pi_{z_i z_j} \right) + log\,(\beta_{m,l}) + \log\,(\theta_i^m) + \log(\theta_j^l) \right\}$$

$$-\theta_i^m \theta_j^l \beta_{m,l} \pi_{z_i z_j} = \sum_{m,l} \sum_{i} k_i^{m,l} \log\,(\theta_i^m) + \sum_{m,l} \sum_{q \leq s} e_{qs}^{m,l} \left\{ \log\,(\pi_{qs}) + log\,(\beta_{m,l}) \right\} - \sum_{q \leq s} \pi_{qs}$$

Maximizes likelihood estimates parameters

$$\widehat{\theta}_i^m = \frac{\sum_{m,l} k_i^{m,l}}{\sum_{m,l} \sum_{i:z_i=q} k_i^{m,l}} = \frac{\sum_{m,l} k_i^{m,l}}{\sum_{m,l} e_q^{m,l}}$$

$$\widehat{\pi}_{qs} = \sum_{m,l} \sum_{i,j:z_i=q,z_j=s} A_{ij}^{m,l} = \sum_{m,l} e_{qs}^{m,l},$$

$$\widehat{\beta}_{m,l} = \frac{\sum_{q \leq s} e_{qs}^{m,l}}{\sum_{m,l} \sum_{q \leq s} e_{qs}^{m,l}} = \frac{L^{m,l}}{L}$$

After ignoring sentences that are not related to tagging, the Modularity function as

$$Q_{\mathrm{SDRMLSBM}} = \sum_{m,l} \sum_{q \leq s} \left\{ e_{qs}^{m,l} \log \left(\frac{\sum_{m,l} e_{qs}^{m,l}}{\left(\sum_{m,l} e_q^{m,l} \right) \left(\sum_{m,l} e_s^{m,l} \right)} \right) \right\}$$

Its normalized version based on each layers

$$Q_{\mathrm{SDRMLSBM}} = \sum_{m,l} \sum_{q \leq s} \left\{ (e_{qs}^{m,l}/N^{m,l}) \log \left(\frac{\sum_{m,l} (e_{qs}^{m,l}/N^{m,l})}{\left(\sum_{m,l} (e_q^{m,l}/N^{m,l}) \right) \left(\sum_{m,l} (e_s^{m,l}/N^{m,l}) \right)} \right) \right\}$$

4. Computations

We use the multilayer version of the algorithm used by [9]. The Kerninghan-Lin algorithm is a graph segmentation algorithm and has a non-greedy approach that leads to more accurate results with given K number of communities. But this algorithm requires start assignment and the final assignment depends on the quality of the initial assignment. This algorithm often gets stuck in the local maximum, so we use several starting points to improve the quality and reach the desired value with the average values of the results. Our algorithm gives the adjacency matrix A as the input, and assigns values $\{1, \ldots, K\}$randomly to the vertices.

The algorithm searches the probability level by moving a vertex from one group r to another group s. After proposing all such moves, on all eligible vertices and communities, it chooses the move that most likely increases. If no improvement is possible, the algorithm selects a motion that minimizes the probability function, as this motion helps to escape the local optimum. We allow each node to move only once, and when all vertices have moved, the states through which the previous system was evaluated and the status with the highest desired score are used as a starting point for repeating the next search. When a complete iteration occurs without any improvement in target amount, algorithm ends.

5. Synthetic Networks Results

In this section, we compare the performance of modularity values to identify communities through a simulation. Because the real assignments of nodes on simulated data are known, we compare the assignment of communities in different models. As a metric, we use Normalized Mutual Information (NMI), a measurement based on information theory that measures similarity between two clusters. This metric shows values between 0 and 1, 0 indicates that the assignment according to the real assignment is random, and 1 indicates a complete match with the real assignment and algorithm assignment. Since the measurement is "normalized", clustering methods can also be compared to a large number of clusters. Finally, assuming that the number of clusters is already known, we perform the clustering accuracy by comparing the values of NMI. All results reported during the experiment are averaged over 100 simulations.

5.1. Data generation

In this section, we propose a methodology for generating heterogeneous multilayer networks. The α parameter regulates the amount of crossing of communities between inter-layers. This network contains a number of different M layers in which each $L_i$layer has $N_i$nodes With density $d_i = density(L_i)$. The method consists of the following three steps:

Step 1. Intra-layer Networks Block Matrix:. We generate data with a degree corrected multi-layered stochastic block model. First label the N_i nodes with $|C_i|$ communities. We consider the community. The size of network communities can be balanced (of equal size) or unbalanced (of unequal size). Then we generate a different connection matrix for each layer using a stochastic block model. In our stochastic block model, the connection matrix is considered more probability values for intra-block edges than for inter-block edges. To control the strength signal of connection matrix values of the random connection matrix in diameter adds amounts that subtracted from non-diameter values. ω_{rs}^{random} represents a completely random network whose values are the expected values in a random graph with constant degrees, which we have $\omega_{rs}^{random} = \kappa_r\kappa_s/2m$. Where $\kappa_r = n_s d_i$that $n_s$is the number of nodes in the community s and $d_i$is the density of that layer. This means that by subtracting from the non-diameter elements of the matrix ω_{rs}^{random} and adding it to the elements of the diameter. For a strong signal, the diameter entries are 3 to 4 times more than entries in that row or column. For a weak signal, the diameter entries are equal to entries in that row or column.

$$\omega^{random} = \begin{bmatrix} n_1 n_1 d & n_1 n_2 d & \cdots \\ n_1 n_2 d & \ddots & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

By adding and subtracting, entries on the diameter and non-diameter can determine signal.

$$\begin{bmatrix} n_1 n_1 d + x & n_1 n_2 d - x & \cdots \\ n_1 n_2 d - x & \ddots & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

The degree parameter is generated using the power law distribution. For each node of each layer in the power law distribution, we consider an independent parameter.

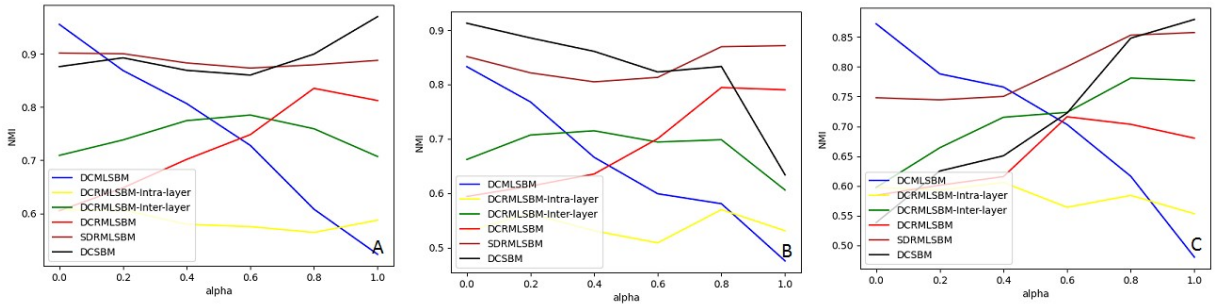

Figure 3: Simulation of strong intra layers and strong inter layers and the alpha parameter is variable (A) intra-layer-density =0.2 and inter-layer-density = 0.2 (B) intra-layer-density =0.2 and inter-layer-density = 0.05 (C) intra-layer-density =0.05 and inter-layer-density = 0.2

Step 2. Inter Layer Communities:. We combine the community $x_i = C_i$from layer $L_i$with the community $x_j = C_j$from layer $L_j$to generate the community x_{ij}. Assuming $|C_i|$ and $|C_j|$ as the number of communities of layers $L_i$and L_j, therefore $|C_c| = min\{|C_i|, |C_j|\}$ specifies the maximum possible number of communities in inter layers. We construct$|C_c| \times \alpha$ communities of inter-layers by crossing the inter layer communities of both $L_i$and L_j randomly.

Step 3. Inter Network Block Matrix: . To build inter networks similar to intra networks, except that here the diameter entry is considered only to the communities that were crossed in the previous step, and the communities that are not crossed on the diameter are zero. For crossed communities, diameters entries must be determined similar first step.

For example, if only community 1 was crossed in the previous step, the matrix ω^{random} as

$$\omega^{random} = \begin{bmatrix} n_1n_1d & n_1n_2d & n_1n_3d \\ n_2n_1d & 0 & n_2n_3d \\ n_3n_3d & n_3n_2d & 0 \end{bmatrix}$$

By adding and subtracting, entries on the diameter and non-diameter can determine signal.

$$\begin{bmatrix} n_1n_1d + x + y & n_1n_2d - x & n_1n_3d - y \\ n_2n_1d - x & 0 & n_2n_3d \\ n_3n_1d - y & n_3n_2d & 0 \end{bmatrix}$$

5.2. Simulation Results

For this simulation, we consider the value of M to be 3 and $N_i$to be equal to each other at 100, and the $C_i$to be equal to 5. We also consider the size of communities to be balanced. According to the strong signal and weak of the layers, three modes can be considered for the signal of layers:

(1) Strong intra layers and strong inter layers
(2) Strong intra layers and weak inter layers
(3) Weak intra layers and strong inter layers

The NMI criterion must be examined according to the change in the density of the layers and the α parameter.

5.2.1. Strong intra layers and strong inter layers

To investigate the effect of density on the performance of our models, the α parameter must change and we consider three different states: in fact, three different states of density of intra layers and inter layers, so that the intra layers are high and the inter layers are high. Intra layers are high and inter layers are low. Intra layers are low and inter layers are high. The results shows in Figure 3.

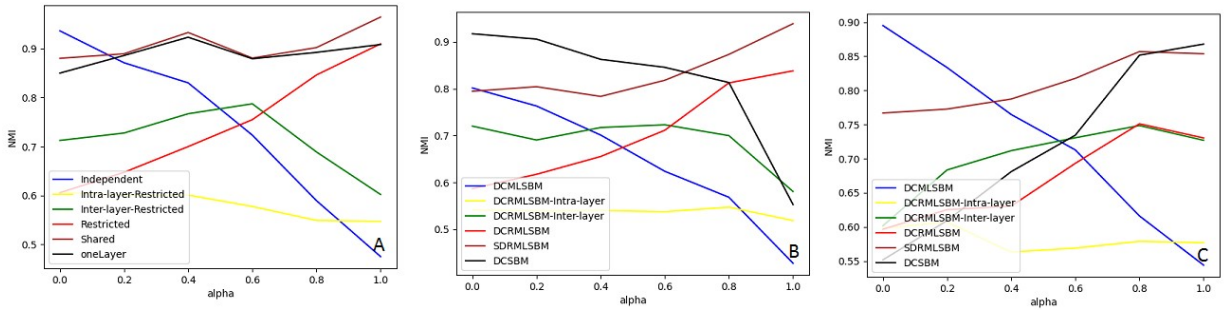

Figure 4: Simulation of strong intra layers and weak inter layers and the alpha parameter is variable (A) intra-layer-density =0.2 and inter-layer-density = 0.2 (B) intra-layer-density =0.2 and inter-layer-density = 0.05 (C) intra-layer-density =0.05 and inter-layer-density = 0.2

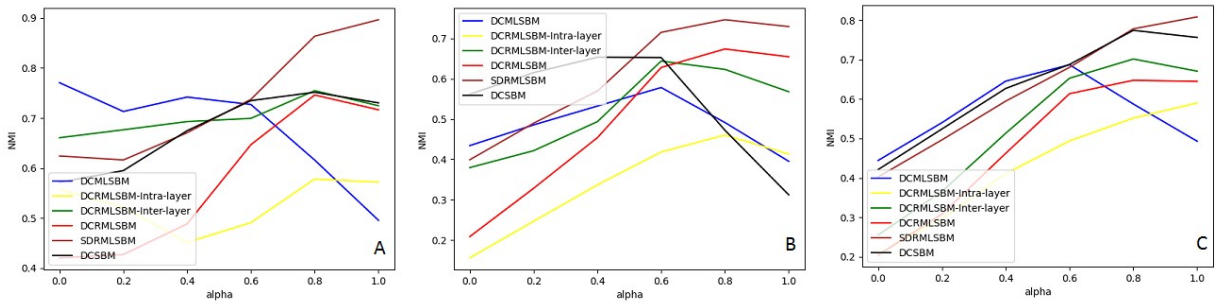

Figure 5: Simulation of weak intra layers and strong inter layers and the alpha parameter is variable (A) intra-layer-density =0.2 and inter-layer-density = 0.2 (B) intra-layer-density =0.2 and inter-layer-density = 0.05 (C) intra-layer-density =0.05 and inter-layer-density = 0.2

5.2.2. Strong intra layers and weak inter layers

To investigate the effect of density on the performance of our models, the α parameter must change and we consider three different states: in fact, three different states of density of intra layers and inter layers, so that the intra layers are high and the inter layers are high. Intra layers are high and inter layers are low. Intra layers are low and inter layers are high. The results shows in Figure 4.

5.2.3. Weak intra layers and strong inter layers

To investigate the effect of density on the performance of our models, the α parameter must change and we consider three different states: in fact, three different states of density of intra layers and inter layers, so that the intra layers are high and the inter layers are high. Intra layers are high and inter layers are low. Intra layers are low and inter layers are high. The results shows in Figure 5. As seen in different simulations. The DCMLSBM performance is better in the smaller amount of the alpha parameter, and The DCRMLSBM performance is better in the larger amount of the alpha parameter. It seems that the reason is also in the crossing of communities between different layers.

However, in the states that intra-layer-density are high and inter-layer-density are low and the α parameter is low, one layer model has better performance in comparison to DCMLSBM.

6. Real Networks Results

6.1. DBLP Dataset

A subset of the DBLP dataset which is a computer science bibliography website was extracted in (Gao et al., 2009). DBLP has more than 3.4 million journal articles, conference papers, and other publications in computer science.

Table 1: The NMI of clustering from different community detection methods for DBLP and AMiner datasets

	DCMLSBM	DCRMLSBM-Intra-Layer	DCRMLSBM-Inter-Layer	DCRMLSBM	SDRMLSBM	DCSBM
DBLP	0.920014	0.920014	0.882668	0.882668	0.885119	0.85599
AMiner	0.901216	0.901216	0.872698	0.872698	0.875139	0.84691

This subset has bibliographical records from four research areas: database, data mining, information retrieval, and artificial intelligence that are its labels for detecting. This network contains three types of nodes: paper, conference, and author and two types of edges: paper-conference (paper published at conference), paper-author (paper written by author). This dataset has 14,376 papers written by 14,475 authors, and published at 20 conferences that are labeled with the research areas. Just only 4,057 authors have true research area that are connected to a subset of 14,328 papers, covering all 20 conferences. Finding the true research areas of the authors is our problem. We apply our data analysis algorithms on this labeled subset of the data because error rates can be computed on labeled data.

Applying our five different proposed methods and one layer DCSBM for 100 reputations. The results shows in Table 1.

6.2. AMiner Dataset

AMiner uses social network analysis to search and perform data mining operations against academic publications on the Internet and identifying connections between researchers, conferences, and publications. It has indexed 130,000,000 researchers and more than 265 million publications.

A subset of the AMiner was extracted that contains three types of nodes: paper, conference, and author and two types of edges: paper-conference (paper published at conference), paper-author (paper written by author). This dataset has 127,623 papers written by 164,472 authors, and published at 101 conferences that are labeled with 10 research areas. . Just only 127,202 papers have true research area. Finding the true research areas of the authors is our problem. We apply our data analysis algorithms on this labeled subset of the data because error rates can be computed on labeled data.

Applying our five different proposed methods and one layer DCSBM for 100 reputations. The results shows in Table 1.

7. Conclusions

In this paper, we have described stochastic block models for heterogeneous networks. These models can be divided into two broad categories, based on independent degree and those based on common degree. While independent degree models have a separate degree parameter in each layer for each node, shared degree base models share degree information in layers. Independent degree model has better performance at more local communities networks (lower alpha parameter) in contrast shared degree model has better performance at more global communities networks (higher alpha parameter). But there is exception in low inter layers density networks that one layer degree corrected stochastic block model has better performance.

References

D. Liu, L. L., Ma, Z., 2020. A community detection algorithm for heterogeneous information networks. IEEE Access 8, 195655–195663.

Gao, J., Liang, F., Fan, W., Sun, Y., Han, J., 2009. Graph-based consensus maximization among multiple supervised and unsupervised models. Advances in Neural Information Processing Systems 22, 585–593.

Han, Q., Xu, K., Airoldi, E., 2015. Consistent estimation of dynamic and multilayer block models. 32nd International Conference on Machine Learning, 1511–1520.

Karrer, B., Newman, M. E. J., 2011. Stochastic block models and community structure in networks. Phys. Rev. E. 83.

Larremore, D. B., Clauset, A., Jacobs, A. Z., 2014. Efficiently inferring community structure in bipartite networks. Physical Review E 90.

Li, Z., Pan, Z., Zhang, Y., Li, G., Hu, G., 2016. Efficient community detection in heterogeneous social networks. Hindawi Publishing Corporation Mathematical Problems in Engineering.

Paul, S., Chen, Y., 2015. Community detection in multi-relational data with restricted multi-layer stochastic blockmodel. arXiv:1506.02699.

Paul, S., Chen, Y., 2016. Null models and modularity based community detection in multi-layer networks. arXiv:1608.00623.

Peixoto, T. P., 2015. Inferring the mesoscale structure of layered, edge-valued, and timevarying networks. Physical Review E 92.
Pramanik, S., Tackx, R., Navelkar, A., Guillaume, J. L., Mitra, B., 2017. Discovering community structure in multilayer networks. International Conference on Data Science and Advanced Analytics, 611–620.
Sengupta, S., Chen, Y., 2015. Spectral clustering in heterogeneous networks. Statistica Sinica 15, 1081–1106.
Song, J., Tang, S., Liu, X., Gao, Y., Yang, H., Lu, P., 2015. . a modularity-based method reveals mixed modules from chemical-gene heterogeneous network. PLOS ONE 10, 1–16.
Stanley, N., Shai, S., Taylor, D., Mucha, P. J., 2015. Clustering network layers with the strata multilayer stochastic block model. arXiv:1507.01826.
Sun, Y., Han, 2021. Mining heterogeneous information networks: principles and methodologies. Synthesis Lectures on Data Mining and Knowledge Discovery, 1–159.
Taylor, D., Shai, S., Stanley, N., Mucha, P. J., 2016. Enhanced detectability of community structure in multilayer networks through layer aggregation. Physical Review Letters 116, 228–301.
Valles-Catala, T., Massucci, F., Guimera, R., Sales-Pardo, M., 2016. Multilayer stochastic block models reveal the multilayer structure of complex networks. Physical Review X 6.
X. Liu, W. Liu, T. M., Wakita, K., 2014. A framework for community detection in heterogeneous multi-relational networks. Advances in Complex Systems, 20–30.
Zhang, J., Chen, Y., 2018. Modularity based community detection in heterogeneous networks. arXiv:1803.07961.