Community Detection for Multilayer Heterogeneous Network

Fan Yang and Fengshuo Zhang

Department of Statistics, University of Chicago

May 18, 2017

Abstract

Many real world networks consist of multiple types of nodes with edges that are heterogeneous in nature. However, most of the existing work for community detection only focused on homogeneous network consisting of a single layer. In this paper, we propose a modified Degree-Corrected Stochastic Model (DCBM) for modeling multilayer heterogeneous network. We develop a spectral clustering method that can unify the information contained in each sub-network, and demonstrate its efficiency to detect communities on simulated data and on Authorship/Citation network data. As a by-product, we present a novel algorithm called BiScore for clustering bipartite network under DCBM, and show that under mild conditions BiScore is guaranteed to yield consistent results.

1 Introduction

Network data are pervasive nowadays and an important problem is to identify the community structure of the network. State-of-the-art methods for community detection can be roughly divided into two classes: algorithm-based approaches and model-based approaches, where the latter are built on certain generative models for networks. Degree-Corrected Stochastic Block Model (DCBM) [13], an extension of Stochastic Block Model (SBM) [8], is a popular network model that has attracted a lot of attention recently [12, 21, 27, 6]. Compared to SBM, DCBM allows for degree heterogeneity and hence is considered to be a better choice for modelling many real world networks.

Although community detection for networks has been extensively studied in the literature, most of the existing work only focused on homogeneous network consisting of a single layer. However, in many real world examples the network of interest often contain multiple layers, where nodes in different layers are of different types and edges in different sub-network are heterogeneous in nature. For instance, Authorship/Citation network [11, 3, 18] can be regarded as a two layer network: one layer consists of authors where edges between two nodes represent the friendship between two authors, and the other layer consists of papers where edges between two nodes imply the citations between two papers; moreover, edges between two layers indicate which author wrote which paper. Another example is the Drug-Protein Interaction network [2, 26, 25], where proteins and drugs constitute their own layer respectively; edges between two layers indicate the interactions between proteins and drugs, whereas edges within a layer represent certain relationship between nodes, like the presence of side-effect of two drugs, structural or functional similarity of two proteins, etc. Community detection is often required for each layer of such network, and information is inevitably lost if we simply consider each layer separately without incorporating the data across layers.

In this article, we propose a modified DCBM called Multi-DCBM for multilayer heterogeneous network. Multi-DCBM is essentially a collection of (classical) DCBM for each sub-network with the key assumption that nodes in each layer share the same community structure across different types of edges. In order to accurately recover the community structure in our model, we have to utilize the information contained in every bipartite network. To this end, we develop a new method called BiScore, a variant of Score proposed in [12], to cluster nodes for bipartite network under DCBM. Under mild conditions, we prove that BiScore is (weakly) consistent. Moreover, we propose the MultiScore algorithm that combines the strength of Score and BiScore to simultaneously detect communities for each layer in Multi-DCBM. Numerical studies demonstrate the efficiency of these methods and an application on Authorship/Citation network data leads to interesting results.

Connection to Existing Literature: The study of heterogeneous network is relatively new and has begun receiving attention from researchers especially from computer science community [24]. Existing community detection methods for heterogeneous network include modularity optimization [17] and non-negative matrix factorization [15]. Among these works, [22] adopts a model-based approach and is most similar to the current work. In [22], each community is allowed to contain nodes from several types and the authors employ the classical DCBM in their study; whereas in our model each community is limited to a single layer containing only one type of nodes and a modified DCBM is assumed where every node has multiple degree parameters. Our clustering method is different from [22] as well.

Some researchers have studied multilayer network where every layer shares the same set of nodes but have different sets of edges [4, 10, 19]. For community detection purpose, a natural extension of SBM is commonly adopted [20, 7]. Our paper differs from this line of work in that the "layers" we considered here are not only edge-heterogeneous but also node-heterogeneous: nodes in different layers are distinct and represent different entities.

For bipartite network, it is common to apply modularity-based approaches [16] or variants of maximum likelihood [14] to cluster nodes. As far as we know, our BiScore algorithm is the first spectral clustering method designed for bipartite network that provably yields consistent community detection under DCBM and hence is of independent interest.

2 Model and Methods

2.1 Multi-DCBM for Multilayer Heterogeneous Network

Suppose we observe an undirected network $G = (V, E)$ [1] where the nodes are of Q different types

$$V = V^1 \cup V^2 \cup \cdots \cup V^Q$$

We call each V^i a *layer* of G and there are Q layers in total. Let A be the adjacency matrix of G and we can write A into blocks $A = (A^{i,j})_{1\le i,j\le Q}$, where $A^{i,j}$ is a $|V^i| \times |V^j|$ adjacency matrix representing the sub-network between layer i and layer j: $A^{i,j}(p,q) = 1$ if there is an edge between pth node in V^i and qth node in V^j; $A^{i,j}(p,q) = 0$ otherwise. Here we allow $A^{i,j}$ to be entirely

[1] For simplicity, we assume the network is undirected in this paper. Our model and methods can be easily generalized to the directed case.

missing for some i and j. In the Authorship/Citation network example, there are $Q = 2$ layers; $A^{1,1}$ is the adjacency matrix of authors' friendship network, $A^{2,2}$ is the adjacency matrix of papers' citation network, and $A^{1,2}$ is the adjacency matrix of the bipartite network representing which author wrote which paper. In practice, the data of friendship among authors may not be accessible, in which case $A^{1,1}$ is missing entirely.

Suppose layer i splits into K_i different communities for $1 \le i \le Q$ [2], that is,

$$V^i = V^i_{(1)} \cup \cdots \cup V^i_{(K_i)} \qquad i = 1, \cdots, Q$$

and we would like to detect communities for all V^i simultaneously given adjacency matrix A. We assume

$$A^{k,l} = \mathbb{E}[A^{k,l}] + W^{k,l}, \qquad \mathbb{E}[A^{k,l}] = \Omega^{k,l} - \mathbb{I}_{k=l}\operatorname{diag}(\Omega^{k,l})$$

for $1 \le k, l \le Q$, where $W^{k,l} = A^{k,l} - \mathbb{E}[A^{k,l}]$ is a generalized Wigner matrix.

It is possible to directly incorporate DCBM here by essentially ignoring layer partitions. In DCBM, for each pair k, l there is a $K_k \times K_l$ community core matrix $P^{k,l}$ such that

$$\Omega^{k,l}(i,j) = \theta_i \theta_j P^{k,l}(g_i, g_j)$$

where the ith node in V^k belongs to $V^k_{(g_i)}$ and the jth node in V^l belongs to $V^l_{(g_j)}$. Here θ_i is a node-specific parameter called *degree heterogeneous parameter*.

Although classical DCBM is an option, we argue that it is not an appropriate model for the multilayer heterogeneous network. This is because DCBM assigns a *single* degree heterogeneous parameter for each node, which does not account for the fact that the propensity of a node to connect to nodes from different layers can be very different. In the Authorship/Citation network example, it is possible that an author is productive but is not social at all or vice versa, thus it is reasonable to assign two degree heterogeneous parameters for each author where one parameter is employed in authors' friendship sub-network and the other one is utilized in author-paper sub-network. Following this line of reasoning, we propose a modified DCBM called Multi-DCBM where each node i is associated with Q degree heterogeneous parameters $\theta_i^{(1)}, \cdots, \theta_i^{(Q)}$; moreover,

$$\Omega^{k,l}(i,j) = \theta_i^{(l)} \theta_j^{(k)} P^{k,l}(g_i, g_j)$$

Multi-DCBM is essentially a collection of DCBMs for each sub-network $A^{i,j}$, $1 \le i, j \le Q$ with the assumption that each layer maintains the same community structure across different sub-networks. It is critical to make such assumptions, as otherwise it would be impossible to borrow strength from heterogeneous networks to perform clustering.

2.2 Score and BiScore

We first give an overview of Score proposed in [12]. Score is a spectral clustering method designed for the (classical) DCBM. Suppose A is the adjacency matrix of the network where the n nodes split into K communities. The Score algorithm is given as follows:

- Obtain the K leading eigenvectors of A, say, $\xi_1, \cdots, \xi_K$.
- Let R be the $n \times (K-1)$ matrix containing element-wise ratios between the first leading eigenvectors and each of the other ones, i.e. $R(i,k) = \xi_{k+1}(i)/\xi_1(i)$, $1 \le i \le n, 1 \le k \le K-1$.

[2]We assume K_i is known for all i throughout the paper

- Apply k-means algorithm on rows of R with K clusters where each row of R is treated as a point in $\mathbb{R}^{K-1}$.
- The ith node is assigned to community j if ith row of R is assigned to the jth cluster.

For bipartite network, Score can not be applied directly as now the adjacency matrix is rectangular. To address this issue, we propose BiScore which employs singular value decomposition (SVD) instead of eigenvalue decomposition. To be specific, suppose we observe the $n_1 \times n_2$ adjacency matrix A of a bipartite network, where V^1 containing nodes corresponding to rows of A splits into K_1 communities and V^2 containing nodes corresponding to columns of A splits into K_2 communities. The BiScore algorithm is given as follows:

- Obtain the $K = \min\{K_1, K_2\}$ leading left singular vectors and K leading right singular vectors of A, say, $u_1, \cdots, u_K$ and $v_1, \cdots, v_K$, respectively.
- Let R_1 be the $n_1 \times (K-1)$ matrix such that $R_1(i,k) = u_{k+1}(i)/u_1(i)$, $1 \le i \le n_1, 1 \le k \le K-1$, and let R_2 be the $n_2 \times (K-1)$ matrix such that $R_2(i,k) = v_{k+1}(i)/v_1(i)$, $1 \le i \le n_2, 1 \le k \le K-1$.
- Apply k-means algorithm on rows of R_1 with K_1 clusters and on rows of R_2 with K_2 clusters.
- The ith node in $V^1(V^2)$ is assigned to community j if ith row of $R_1(R_2)$ is assigned to the jth cluster.

Details of BiScore are presented in section 3, whereas the details of Score can be found in [12].

2.3 MultiScore

In this section, we present MultiScore for multilayer heterogeneous network under Multi-DCBM. The key idea is to combine the information contained in post-ratio singular vectors (or eigenvectors) for each sub-network and apply k-means on rows of the pooled matrix for each layer. In particular, let Obs be the set of indices of observed adjacency matrices, i.e. $\text{Obs} = \{(k,l) | \ A^{k,l} \text{ is observed}\}$. The MultiScore algorithm for community detection on V^q for $q = 1, \cdots, Q$ is given as follows:

- For each adjacency matrix $A^{q,l}$ such that $(q,l) \in \text{Obs}$, obtain the leading $\min\{K_q, K_l\}$ left singular vectors of $A^{q,l}$. Denote them as $u_1^{q,l}, \cdots, u_{\min\{K_q,K_l\}}^{q,l}$.
- Let $R^{q,l}$ be the $|V_q| \times (\min\{K_q, K_l\} - 1)$ matrix such that $R^{q,l}(i,k) = u_{k+1}^{q,l}(i)/u_1^{q,l}(i)$ for $1 \le i \le |V_q|, 1 \le k \le \min\{K_q, K_l\} - 1$.
- Stack matrices $\{w_{q,l} R^{q,l}\}_{l:(q,l)\in\text{Obs}}$ horizontally into a single concatenated matrix R where $w_{q,l}$ are nonnegative weights. Apply k-means on rows of R with K_q clusters.
- The ith node in V^q is assigned to community j if ith row of R is assigned to the jth cluster.

We suggest two ways to choose the weights in MultiScore:

- *Simple Pool.* Let $w_{q,l} = 1$ for all l. This criteria assigns equal weight to every sub-network.
- *Signal Adjusted Pool.* Let $w_{q,l} = \sigma^2_{\min\{K_q,K_l\}}/\sigma_1$ where σ_i is the ith leading singular value of $A^{q,l}$. The weight assigned here is a measures of signal-to-noise ratio of a network [1], so this criteria tends to put more weights on those sub-networks that are presumably more useful.

The weights can also be determined in prior according to user's domain knowledge or belief that some sub-networks are more informative than others.

3 Theoretical Analysis of BiScore

Consider a bipartite undirected network $G = (V, E)$ where nodes splits into two sub-graphs $V = V^1 \cup V^2$ and every edge connects a node in V^1 to one in V^2. Suppose $|V^1| = n_1$ and $|V^2| = n_2$, and we observe the $n_1 \times n_2$ adjacency matrix A where $A_{ij} = 1$ if and only if there is an edge between ith node in V^1 and jth node in V^2. Suppose there are K_1 communities for V^1 and K_2 communities for V^2, i.e. $V^1 = V^1_{(1)} \cup \cdots \cup V^1_{(K_1)}$ and $V^2 = V^2_{(1)} \cup \cdots \cup V^2_{(K_2)}$. In bipartite DCBM, we assume

$$A = \mathbb{E}(A) + W, \qquad \mathbb{E}(A) = \Omega$$

where $W = A - \mathbb{E}(A)$ is a generalized Wigner matrix. We assign the degree heterogeneous parameters $\{\theta(i)\}_{i=1}^{n_1}$ for nodes in V^1 and $\{\gamma(i)\}_{i=1}^{n_2}$ for nodes in V^2. Let P be a $K_1 \times K_2$ matrix and we assume

$$\Omega(i,j) = \theta(i)\gamma(j)P(g_i, g_j) \quad \text{if } i \in V^1_{(g_i)} \text{ and } j \in V^2_{(g_j)}$$

For identifiability, we fix two constants $g_1, g_2 \in (0,1)$ and assume that

$$\max_{1 \le i \le K_1, 1 \le j \le K_2} P(i,j) = 1, \qquad 0 < \theta_{\min} \le \theta_{\max} \le g_1, \quad 0 < \gamma_{\min} \le \gamma_{\max} \le g_2$$

Moreover, we assume P is of rank $K = \min\{K_1, K_2\}$. In the following analysis, we use n_1 and n_2 as the driving asymptotic parameter, and allow θ and γ to depend on n_1 and n_2. However, we keep K_1, K_2 and P as fixed. Our asymptotic setting is analogous to [12].

3.1 Spectral Analysis of Ω

We first characterize the leading singular values and singular vectors of Ω. For $1 \le k \le K_1$, let $\theta^{(k)}$ be the $n_1 \times 1$ vectors such that $\theta^{(k)}(i)$ equals to $\theta(i)$ if $i \in V^1_{(k)}$ and equals to 0 otherwise. Similarly, for $1 \le k \le K_2$ let $\gamma^{(k)}$ be the $n_2 \times 1$ vectors such that $\gamma^{(k)}(i) = \gamma(i)$ equals to $\gamma(i)$ if $i \in V^2_{(k)}$ and equals to 0 otherwise. Let D_1 be the $K_1 \times K_1$ diagonal matrix of the overall degree intensities for θ, and let D_2 be the $K_2 \times K_2$ diagonal matrix of the overall degree intensities for γ

$$D_1(i,i) = \|\theta^{(i)}\| / \|\theta\|, \quad D_2(j,j) = \|\gamma^{(j)}\| / \|\gamma\|, \quad 1 \le i \le K_1, 1 \le j \le K_2$$

The following result is an extension of Lemma 2.1 in [12].

Lemma 1. *Suppose all $K = \min\{K_1, K_2\}$ nonzero singular values of $D_1 P D_2$ are nondegenerate. Let $\frac{\sigma_1}{\|\theta\|\|\gamma\|}, \ldots, \frac{\sigma_K}{\|\theta\|\|\gamma\|}$ be such singular values, and let $a_1, \ldots, a_k$ be the associated (unit norm) right singular vectors and $b_1, \ldots, b_K$ be the associated (unit norm) left singular vectors. Then the K nonzero singular values of Ω are $\sigma_1, \ldots, \sigma_K$ with the associated (unit-norm) left and right singular vectors being*

$$u_k = \sum_{i=1}^{K_1} \frac{b_k(i)}{\|\theta^{(i)}\|} \cdot \theta^{(i)}, \quad v_k = \sum_{i=1}^{K_2} \frac{a_k(i)}{\|\gamma^{(i)}\|} \cdot \gamma^{(i)}, \quad k = 1, 2, \ldots, K$$

3.2 Spectral Analysis of A

Now we characterize the leading singular values and singular vectors of A. We assume

$$\frac{\log(n_1+n_2)\max\{\theta_{\max}\|\gamma\|_1,\gamma_{\max}\|\theta\|_1\}}{\|\theta\|^2\|\gamma\|^2}\to 0 \quad \text{as } n_1,n_2\to\infty \tag{1}$$

The following lemma gives a bound on the spectral norm of W.

Lemma 2. *If* (1) *holds, then with probability at least* $1-o((n_1+n_2)^{-3})$,

$$\|W\|\le 3\sqrt{\log(n_1+n_2)\max\{\theta_{\max}\|\gamma\|_1,\gamma_{\max}\|\theta\|_1\}}$$

We need some mild conditions on D_1PD_2 in order to ensure that K leading singular values of A are non-degenerate. We assume that for some positive constant C_1,C_2,

$$\min_{q\le i\le K}\{\sigma_i-\sigma_{i+1}\}\ge C_1,\ \max_{1\le i,j\le K_1}\{\|\theta^{(i)}\|/\|\theta^{(j)}\|\}\le C_2,\ \max_{1\le i,j\le K_2}\{\|\gamma^{(i)}\|/\|\gamma^{(j)}\|\}\le C_2 \tag{2}$$

where σ_i are the ith largest singular values of D_1PD_2. It follows that all nonzero singular values of D_1PD_2 are bounded away from 0 or ∞ by some constant. Combining Lemma 1 and Lemma 2 together with Weyl's inequality for singular values, we get the following lemma, the proof of which is omitted here.

Lemma 3. *Suppose* (1) *and* (2) *hold. Let* $\hat\sigma_1,\hat\sigma_2,\dots,\hat\sigma_K$ *be the leading singular values of* A*, and let* $\frac{\sigma_1}{\|\theta\|\|\gamma\|},\dots,\frac{\sigma_K}{\|\theta\|\|\gamma\|}$ *be the leading singular values of* D_1PD_2*. With probability at least* $1-o((n_1+n_2)^{-3})$*, the* K *leading singular values of* A *are non-degenerate, and*

$$\max_{1\le k\le K}\{|\hat\sigma_k-\sigma_k|\}\le 3\sqrt{\log(n_1+n_2)\max\{\theta_{\max}\|\gamma\|_1,\gamma_{\max}\|\theta\|_1\}}$$

Combining Lemma 3 with (1), we know with probability at least $1-o(n^{-3})$ we have $\hat\sigma_k\asymp\|\theta\|\|\gamma\|$ for $1\le k\le K$. The following lemma gives a perturbation bound of singular vectors, which is a direct result of Lemma 2 plus a generalized Davis-Kahan theorem (see [23, 5]). The proof is omitted here.

Lemma 4. *Suppose* (1) *and* (2) *hold. Recall* $u_1,u_2,\dots,u_K$ *and* $v_1,v_2,\dots,v_K$ *are the leading left and right singular vectors of* Ω*. Let* $\hat u_1,\hat u_2,\dots,\hat u_K$ *and* $\hat v_1,\hat v_2,\dots,\hat v_K$ *be the leading left and right singular vectors of* A*. With probability at least* $1-o((n_1+n_2)^{-3})$*, for all* $1\le k\le K$*,*

$$\max\{\|u_k-\hat u_k\|^2,\|v_k-\hat v_k\|^2\}\le C\frac{\log(n_1+n_2)\max\{\theta_{\max}\|\gamma\|_1,\gamma_{\max}\|\theta\|_1\}}{\|\theta\|^2\|\gamma\|^2}$$

3.3 Properties of BiScore

In section 2.2 we introduced BiScore for community detection in bipartite network. The following version of BiScore is slightly different from the one presented before:

- Obtain the $K=\min\{K_1,K_2\}$ (unit norm) leading left and right singular vectors of A: $\hat u_1,\hat u_2,\dots,\hat u_K$ and $\hat v_1,\hat v_2,\dots,\hat v_K$.

- Fixing a threshold T_{n_1,n_2}, define $n_1 \times (K-1)$ matrix $\hat{R}_1$ for V^1 such that for $1 \le i \le n_1$ and $1 \le k \le K-1$

$$\hat{R}_1(i,k) = \begin{cases} \hat{u}_{k+1}(i)/\hat{u}_1(i), & \text{if } |\hat{u}_{k+1}(i)/\hat{u}_1(i)| \le T_{n_1,n_2} \\ T_{n_1,n_2}, & \text{if } \hat{u}_{k+1}(i)/\hat{u}_1(i) > T_{n_1,n_2} \\ -T_{n_1,n_2}, & \text{if } \hat{u}_{k+1}(i)/\hat{u}_1(i) < -T_{n_1,n_2} \end{cases}$$

  Similarly, we can define $n_2 \times (K-1)$ matrix $\hat{R}_2$ for V^2 using $\hat{v}$ instead of $\hat{u}$.

- Apply k-means algorithm on rows of $\hat{R}_1$ with K_1 clusters and on rows of $\hat{R}_2$ with K_2 clusters.

- The ith node in $V^1(V^2)$ is assigned to community j if ith row of $R_1(R_2)$ is assigned to the jth cluster.

Here the thresholding procedure is imposed mainly for technical reasons and can be omitted in practice. For convenience we take $T_{n_1,n_2} = \log(n_1+n_2)$ in the paper.

The following lemma shows that under mild conditions all coordinates of $\hat{u}_1$ and $\hat{v}_1$ are guaranteed to be positive. It is a direct result of Perron's theorem [9] on A^TA and AA^T, and the proof is omitted here.

Lemma 5. *Recall $\hat{u}_1$ and $\hat{v}_1$ are leading left and right singular vectors of A. If the bipartite network G is connected, then all coordinates of $\hat{u}_1$ and $\hat{v}_1$ are strictly positive.*

We now define $n_1 \times (K-1)$ matrix R_1 and $n_2 \times (K-1)$ matrix R_2 , which are nonstochastic counterparts of $\hat{R}_1$ and $\hat{R}_2$, respectively.

$$R_1(i,k) = u_{k+1}(i)/u_1(i), \quad R_2(j,k) = v_{k+1}(j)/v_1(j) \qquad 1 \le k \le K-1,\ 1 \le i \le n_1,\ 1 \le j \le n_2$$

By Lemma 1, it's easy to see that for all i,j,k, $|R_1(i,k)| \le C$ and $|R_2(j,k)| \le C$ for some constant C. The following theorem gives an upper bound on $\|\hat{R}_1 - R_1\|_F^2$ as well as $\|\hat{R}_2 - R_2\|_F^2$, and is a key result in characterizing the behaviour of BiScore.

Theorem 1. *Suppose* (1) *and* (2) *hold. If $T_{n_1,n_2} = \log(n_1+n_2)$, then with probability at least $1 - o((n_1+n_2)^{-3})$, we have*

$$\|\hat{R}_1 - R_1\|_F^2 \le C\frac{\log(n_1+n_2)^3\kappa}{\theta_{\min}^2\|\gamma\|^2}, \quad \|\hat{R}_2 - R_2\|_F^2 \le C\frac{\log(n_1+n_2)^3\kappa}{\gamma_{\min}^2\|\theta\|^2}$$

where $\kappa = \max\{\theta_{\max}\|\gamma\|_1, \gamma_{\max}\|\theta\|_1\}$.

To measure the performance of BiScore, we evaluate the Hamming error as defined in [12]. Recall $V^1 = V^1_{(1)} \cup \cdots \cup V^1_{(K_1)}$ is the true community partition for V^1. We introduce a $n_1 \times 1$ vector l where $l(i) = k$ if $i \in V^1_{(k)}$. For the clustering result on V^1 given by BiScore, we introduce a $n_1 \times 1$ vector $\hat{l}$ where $\hat{l}(i) = k$ if i is assigned to the kth cluster. Then the expected number of mismatched labels is $H(\hat{l},l) = \sum_{i=1}^n P(\hat{l}(i) \ne l(i))$. Note that the error should not depend on the way we label the cluster, and for this reason the Hamming error is defined as

$$\text{Hamm}(\hat{l},l) = \min_{\pi \in S_{K_1}} H(\hat{l}, l \circ \pi)$$

where π is a permutation of set $\{1,2,\dots,K_1\}$ and S_{K_1} is the collection of all such permutations. Similarly, we can define the Hamming error for clusters on V^2. Now we are ready to state the main theorem as follows. The proof is similar to the proof of Theorem 2.2 in [12] so we omit it here.

Theorem 2. *Suppose* (1) *and* (2) *hold. Let* $\kappa = \max\{\theta_{\max}\|\gamma\|_1, \gamma_{\max}\|\theta\|_1\}$. *Suppose as* n_1, n_2 *goes to infinity*

$$\frac{\log(n_1+n_2)^3\kappa}{\theta_{\min}^2\|\gamma\|^2\min\{|V^1_{(1)}|,|V^1_{(2)}|,\ldots,|V^1_{(K_1)}|\}} \to 0, \quad \frac{\log(n_1+n_2)^3\kappa}{\gamma_{\min}^2\|\theta\|^2\min\{|V^2_{(1)}|,|V^2_{(2)}|,\ldots,|V^2_{(K_2)}|\}} \to 0$$

Then for the estimated label vector $\hat{l}_1$ *on* V^1 *and* $\hat{l}_2$ *on* V^2, *we have for sufficiently large* n_1, n_2,

$$Hamm(\hat{l}_1, l_1) \le C\frac{\log(n_1+n_2)^3\kappa}{\theta_{\min}^2\|\gamma\|^2}, \quad Hamm(\hat{l}_2, l_2) \le C\frac{\log(n_1+n_2)^3\kappa}{\gamma_{\min}^2\|\theta\|^2}$$

where l_1 *and* l_2 *are true label on* V^1 *and* V^2 *respectively.*

Theorem 2 suggests that BiScore is weakly consistent under mild conditions, say, $n_1 = o(\exp(n_2)), n_2 = o(\exp(n_1))$ and $\theta_{\min}, \gamma_{\min} \ge c$ for some constant $c > 0$. Under similar conditions, the consistency of MultiScore follows directly from Theorem 1 and Theorem 2 together with Theorem 2.1 and 2.2 in [12].

4 Experiments

4.1 Simulated Data

We consider a network with two layers V^1 and V^2, where layer V^1 has $N_1 = 600$ nodes and V^2 has $N_2 = 900$ nodes. Each layer has $K = 3$ different communities with equal size. We take the community detection for layer V^1 as an example. Let $A^{1,1} \in \mathbb{R}^{N_1\times N_1}$ denote the adjacency matrix within layer V^1 and $A^{1,2} \in \mathbb{R}^{N_1\times N_2}$ denote the adjacency matrix between V^1 and V^2. We compare four different methods to cluster nodes in V^1: (1) Score on $A^{1,1}$, (2) BiScore on $A^{1,2}$, (3) MultiScore with simple pool on $A^{1,1}$ and $A^{1,2}$, and (4) MultiScore with signal adjusted pool on $A^{1,1}$ and $A^{1,2}$. The network is generated from Multi-DCBM.

For each layer, we fix 3×3 matrix P (from Multi-DCBM) such that $P(i,i) = 1$ and $P(i,j) = 0.5$ for $i \ne j$. The degree heterogeneous parameters are randomly drawn from certain distribution: for $i \in V^1$, $\theta_i^{(1)} \sim 1/\text{Unif}(1,a)$ and $\theta_i^{(2)} \sim 1/\text{Unif}(1,b)$; for $j \in V^2$, $\theta_j^{(1)} \sim 1/\text{Unif}(2,b)$. Here the parameter a and b controls the sparsity of $A^{1,1}$ and $A^{1,2}$, where larger a and b suggests sparser network. In the experiment, we set $b \in \{3,6,10\}$ and vary a from 1 to 10. Figure 1 plots a against averaged misclassification rate (scaled Hamming error) of the four methods over 50 sampled networks for each b. The experiment shows that MultiScore enjoys the smallest misclassification rate compared to the Score and BiScore applied on a single sub-network. Among two versions of MultiScore, the signal adjusted pool is slightly better than the simple pool especially when the network is sparse. These observations indicate that MultiScore can effectively incorporate information in heterogeneous networks as we expected.

4.2 Authorship/Citation Network Data

We analyse the Authorship/Citation Network Data for statisticians. The data were collected by Ji and Jin[11] and were based on all published papers in Annals of Statistics, Biometrika, JASA, and JRSS-B from 2003 to 2012. There are 3607 authors and 3248 papers in total. We construct a two-layer network consisting of the author-paper bipartite network and paper citation network[3],

[3]For simplicity we view citation network as a undirected graph

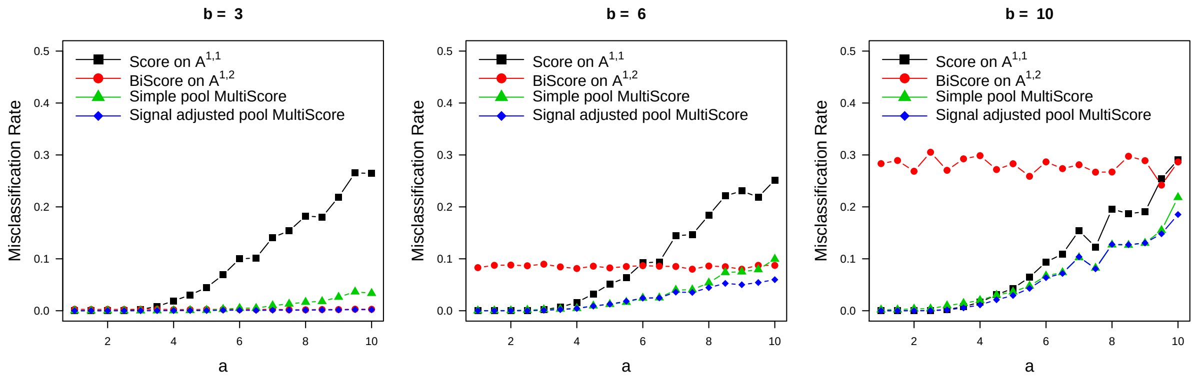

Figure 1: Comparison of Performance for Score, BiScore and two versions of MultiScore under different degree heterogeneity. Larger a and b corresponds to smaller degree heterogeneity.

and following [11] we extract the giant component [4] which contains 1388 authors and 1460 papers. MultiScore was applied to the data set with $K_1 = K_2 = 3$ which results in 3 communities of size $344, 558, 486$ for authors and 3 communities of size $395, 507, 558$ for papers.

We present the results by MultiScore in Figure 2, where for the purpose of visualization we only include papers whose degree are greater than 12 in the citation network (of giant component) and authors who have written at least two of these papers. Table 2 provides the details of papers labelled in Figure 2. Table 1 summarize the basic information of each detected community. By examining the key words of these papers, we find that the three paper community can be interpreted as "Bayesian Statistics", "Nonparametric Methods" and "Variable Selection" respectively. Moreover, each author community exhibits a strong connection with a certain paper community, indicating the active research area of these authors. These results shed light on the research interests of statisticians and the trend of popular research areas in statistics.

Node type	Community(Size)	Representatives / Key words
Authors	□ (344)	Jianqing Fan, Elizaveta Levina, Hui Zou
	□ (558)	Alan Gelfand, David Dunson, John Storey
	□ (486)	Peter Hall, Fang Yao, Iain Johnstone
Paper	○ (395)	Selection(40), Semiparametric(35), Likelihood(30)
	○ (507)	Bayesian(66), Nonparametric(44), Semiparametric(44)
	○ (558)	Nonparametric(56), Functional(38), High-dimensional(33)

Table 1: Summary of the detected communities by MultiScore. The three most frequent key words for papers are presented together with their frequency for each community.

[4] Largest connected subgraph

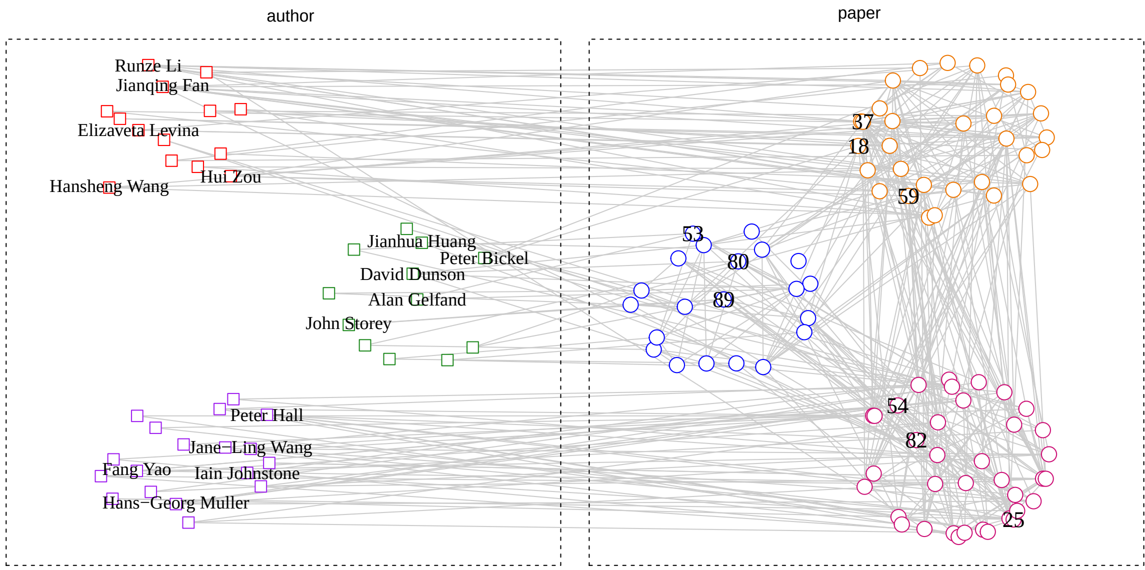

Figure 2: Visualization of community detection results by MultiScore. Only papers whose degree are greater than 12 in the giant component of citation network and authors who have written at least two of these papers are included. Names of five most productive authors (confined to the data set) are shown for each author community, and three representative papers are labelled for each paper community. Details of the labelled paper can be found in Table 2

5 Conclusion

In this paper, we study the community detection problem for multilayer heterogeneous network under a modified degree-corrected stochastic block model. We develop a novel spectral clustering method that can combine the information contained in each sub-networks and demonstrate its efficiency empirically. As a by-product, we propose and analyse a counterpart of Score algorithm for bipartite network. Our framework and tools hence provide a unified way to detect communities when multiple networks containing heterogeneous edges are present.

A Proofs

A.1 Proof of Lemma 1

Let σ_k be the k-th largest singular values of Ω, and denote its corresponding left singular vectors as u_k and right singular vectors as v_k, for $1 \le k \le K$. Let a_k be the $K_2 \times 1$ vector such that $a_k(j) = (\frac{\gamma^{(j)}}{\|\gamma^{(j)}\|}, v_k)$ and let b_k be the $K_1 \times 1$ vector such that $b_k(i) = (\frac{\theta^{(i)}}{\|\theta^{(i)}\|}, u_k)$. We can write Ω as

$$\Omega = \|\theta\|\|\gamma\| \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} (D_1 P D_2)(i,j) \left(\frac{\theta^{(i)}}{\|\theta^{(i)}\|} \right) \left(\frac{\gamma^{(i)}}{\|\gamma^{(i)}\|} \right)^T$$

Index	Title	Authors
54	Functional data analysis for sparse longitudinal data	Fang Yao, Hans-Georg Müller and Jane-Ling Wang
25	Adapting to unknown sparsity by controlling the false discovery rate	Felix Abramovich, Yoav Benjamini, David Donoho, and Iain Johnstone
82	On properties of functional principal components analysis	Peter Hall and Mohammad Hosseini-Nasab
18	Regularized estimation of large covariance matrices	Peter Bickel and Elizaveta Levina
59	The adaptive lasso and its oracle properties	Hui Zou
37	Nonconcave penalized likelihood with a diverging number of parameters	Jianqing Fan and Heng Peng
53	Bayesian nonparametric spatial modeling with Dirichlet process mixing	Alan Gelfand, Athanasios Kottas and Steven MacEachern
80	Bayesian density regression	David B. Dunson and Natesh Pillai
89	Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach	John Storey, Jonathan Taylor and David Siegmund

Table 2: Titles and authors of labelled papers appeared in Figure 2.

and by simple calculations

$$\Omega v_k = \|\theta\|\|\gamma\| \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} (D_1 P D_2)(i,j) a_k(j) \frac{\theta^{(i)}}{\|\theta^{(i)}\|}$$

and

$$\Omega^T u_k = \|\theta\|\|\gamma\| \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} (D_1 P D_2)(i,j) b_k(i) \frac{\gamma^{(j)}}{\|\gamma^{(j)}\|}$$

Since $\Omega^T u_k = \sigma_k v_k$ and $\{\frac{\gamma^{(j)}}{\|\gamma^{(j)}\|}\}_{j=1}^{K_2}$ is an orthonormal basis, we have

$$a_k(j) = (\frac{\gamma^{(j)}}{\|\gamma^{(j)}\|}, \frac{1}{\sigma_k}\Omega^T u_k) = \frac{\|\theta\|\|\gamma\|}{\sigma_k} \sum_{i=1}^{K_1} (D_1 P D_2)(i,j) b_k(i) = \frac{\|\theta\|\|\gamma\|}{\sigma_k} (b_k^T D_1 P D_2)(j)$$

which suggests

$$(D_1 P D_2)^T b_k = \frac{\sigma_k}{\|\theta\|\|\gamma\|} a_k$$

Similarly, we have

$$(D_1 P D_2) a_k = \frac{\sigma_k}{\|\theta\|\|\gamma\|} b_k$$

Hence we know b_k and a_k are left and right singular vectors of D_1PD_2 with singular value $\frac{\sigma_k}{\|\theta\|\|\gamma\|}$. Moreover, we have

$$u_k = \frac{1}{\sigma_k}\Omega v_k = \frac{\|\theta\|\|\gamma\|}{\sigma_k}\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}(D_1PD_2)(i,j)a_k(j)\frac{\theta^{(i)}}{\|\theta^{(i)}\|}$$
$$= \frac{\|\theta\|\|\gamma\|}{\sigma_k}\sum_{i=1}^{K_1}(D_1PD_2a_k)(i)\frac{\theta^{(i)}}{\|\theta^{(i)}\|} = \sum_{i=1}^{K_1}\frac{\theta^{(i)}}{\|\theta^{(i)}\|}b_k(i)$$

and hence $\|b_k\| = \|u_k\| = 1$. Similarly, we have

$$v_k = \sum_{i=1}^{K_2}\frac{\gamma^{(i)}}{\|\gamma^{(i)}\|}a_k(i)$$

with $\|a_k\| = \|v_k\| = 1$. Proof is done as a_k and b_k are uniquely determined up to a sign change.

A.2 Proof of Lemma 2

Let e_i be the $n_1 \times 1$ vector such that $e_i(j) = 1$ if $i = j$ and 0 otherwise, and let $\tilde{e}_i$ be the $n_2 \times 1$ vector such that $e_i(j) = 1$ if $i = j$ and 0 otherwise. We can write $W = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} Z^{(i,j)}$ where $Z^{(i,j)} = W(i,j)e_i\tilde{e}_j^T$. We know $\mathbb{E}[W^2(i,j)] \le \theta(i)\gamma(j)$, and moreover

$$\left\|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbb{E}[Z^{(i,j)}(Z^{(i,j)})^T]\right\| = \left\|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbb{E}[W(i,j)^2]e_ie_i^T\right\| \le \theta_{\max}\|\gamma\|_1$$

and similarly we have $\|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbb{E}[(Z^{(i,j)})^TZ^{(i,j)}]\| \le \gamma_{\max}\|\theta\|_1$. Now let

$$\sigma^2 = \max\{\|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbb{E}[(Z^{(i,j)})^TZ^{(i,j)}]\|, \|\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbb{E}[Z^{(i,j)}(Z^{(i,j)})^T]\|\}$$

, and we know $\sigma^2 \le \max\{\theta_{\max}\|\gamma\|_1, \gamma_{\max}\|\theta\|_1\}$. Fix $q > 0$, and we apply Theorem B.1 in [12] with $h_0 = 1$ and $t = \sqrt{2q\log(n_1+n_2)\max\{\theta_{\max}\|\gamma\|_1, \gamma_{\max}\|\theta\|_1\}}$,

$$\begin{aligned} &P(\|W\| \ge \sqrt{2q\log(n_1+n_2)\max\{\theta_{\max}\|\gamma\|_1, \gamma_{\max}\|\theta\|_1\}}) \\ \le &(n_1+n_2)\exp\left[\frac{-q\log(n_1+n_2)}{1+(1/3)\sqrt{2q\log(n_1+n_2)/(\max\{\theta_{\max}\|\gamma\|_1, \gamma_{\max}\|\theta\|_1\})}}\right] \end{aligned} \tag{3}$$

According to (1), we have by some simple algebra

$$\log(n_1+n_2)/\max\{\theta_{\max}\|\gamma\|_1, \gamma_{\max}\|\theta\|_1\} \to 0$$

The claim follows by taking $q = 9/2$ in (3).

A.3 Proof of Theorem 1

For any $n \times 1$ vector ξ with strictly positive coordinates, we define the coordinate oscillation $\mathrm{OSC}(\xi) = \max_{1\le i,j\le n}\{\xi(i)/\xi(j)\}$. Let Θ be a $n_1 \times n_1$ diagonal matrix with $\Theta(i,i) = \theta(i)$. By similar argument of Lemma 2.6 in [12], we have

$$\mathrm{OSC}(\Theta^{-1}u_1) \le C \tag{4}$$

for some constant C. This suggests that all $u_1(i)/\theta(i)$ are of the same order, and since $\|u_1\| = 1$ we have $u_1(i)/\theta(i) \asymp 1/\|\theta\|$ for $1 \le i \le n_1$.

Fixing a constant $c_0 \in (0,1)$, we define

$$\hat{S} = \{1 \le i \le n_1 : |\hat{u}_1(i)/u_1(i) - 1| \le c_0\}$$

By Lemma 4, we know there is an event E_{n_1,n_2} with $P(E^c_{n_1,n_2}) = o((n_1+n_2)^{-3})$ such that over E_{n_1,n_2}, for all $1 \le k \le K$,

$$\|\Theta^{-1}(\hat{u}_k - u_k)\|^2 \le \|\Theta^{-1}\|^2\|\hat{u}_k - u_k\|^2 \le C\frac{\log(n_1+n_2)\max\{\theta_{\max}\|\gamma\|_1, \gamma_{\max}\|\theta\|_1\}}{\theta^2_{\min}\|\theta\|^2\|\gamma\|^2}$$

For $i \notin \hat{S}$, we have $|(\hat{u}_1(i) - u_1(i))/\theta(i)| > c_0 u_1(i)/\theta(i) \ge C/\|\theta\|$. Hence we get

$$\|\Theta^{-1}(\hat{u}_k - u_k)\|^2 \ge \sum_{i \notin \hat{S}} |(\hat{u}_1(i) - u_1(i))/\theta(i)|^2 \ge C(n_1 - |\hat{S}|)/\|\theta\|^2$$

which implies that

$$n_1 - |\hat{S}| \le C\frac{\log(n_1+n_2)\max\{\theta_{\max}\|\gamma\|_1, \gamma_{\max}\|\theta\|_1\}}{\theta^2_{\min}\|\gamma\|^2} \tag{5}$$

Now we write $\|\hat{R}_1 - R_1\|_F^2 = U_1 + U_2$, where

$$U_1 = \sum_{i\notin\hat{S}}\sum_{j=1}^{K-1}(\hat{R}_1(i,j) - R_1(i,j))^2, \quad U_2 = \sum_{i\in\hat{S}}\sum_{j=1}^{K-1}(\hat{R}_1(i,j) - R_1(i,j))^2$$

For $i \notin \hat{S}$, we know $\hat{R}_1(i,j) \le T_{n_1,n_2}$ and $|R_1(i,j)| \le C$. Hence by (5) we have

$$U_1 \le CT^2_{n_1,n_2}(n_1 - |\hat{S}|) \le C\frac{\log(n_1+n_2)^3\max\{\theta_{\max}\|\gamma\|_1, \gamma_{\max}\|\theta\|_1\}}{\theta^2_{\min}\|\gamma\|^2} \tag{6}$$

For $i \in \hat{S}$, we know $|\hat{u}_1(i)/u_1(i) - 1| \le c_0$. For $1 \le k \le K-1$, we can write

$$|\hat{R}_1(i,k) - R_1(i,k)| \le |\frac{\hat{u}_{k+1}(i)}{\hat{u}_1(i)} - \frac{u_{k+1}(i)}{u_1(i)}| \le W_1 + W_2$$

where

$$W_1 = \frac{|\hat{u}_{k+1}(i) - u_{k+1}(i)|}{\hat{u}_1(i)}, \qquad W_2 = \frac{|u_{k+1}(i)(u_1(i) - \hat{u}_1(i))|}{\hat{u}_1(i)u_1(i)}$$

It's easy to see that $W_1 \le C|\hat{u}_{k+1}(i) - u_{k+1}(i)|\|\theta\|/\theta_{\min}$ and $W_2 \le C|u_1(i) - \hat{u}_1(i)|\|\theta\|/\theta_{\min}$. Therefore, we have

$$\begin{aligned}
\sum_{i\in\hat{S}}|\hat{R}_1(i,k) - R_1(i,k)|^2 &\le C\frac{\|\theta\|^2}{\theta^2_{\min}}\sum_{i=1}^{n_1}\left(|\hat{u}_{k+1}(i) - u_{k+1}(i)|^2 + |u_1(i) - \hat{u}_1(i)|^2\right) \\
&= C\frac{\|\theta\|^2}{\theta^2_{\min}}(\|\hat{u}_{k+1} - u_{k+1}\|^2 + \|\hat{u}_1 - u_1\|^2) \\
&\le C\frac{\log(n_1+n_2)\max\{\theta_{\max}\|\gamma\|_1, \gamma_{\max}\|\theta\|_1\}}{\theta^2_{\min}\|\gamma\|^2}
\end{aligned}$$

and hence

$$U_2 = \sum_{k=1}^{K-1} \sum_{i \in \hat{S}} |\hat{R}_1(i,k) - R_1(i,k)|^2 \leq C \frac{\log(n_1+n_2) \max\{\theta_{\max}\|\gamma\|_1, \gamma_{\max}\|\theta\|_1\}}{\theta_{\min}^2 \|\gamma\|^2} \tag{7}$$

Combining (6) and (7), we have over the event E_{n_1,n_2}

$$\|\hat{R}_1 - R_1\|_F^2 = U_1 + U_2 \leq C \frac{\log(n_1+n_2)^3 \max\{\theta_{\max}\|\gamma\|_1, \gamma_{\max}\|\theta\|_1\}}{\theta_{\min}^2 \|\gamma\|^2}$$

Similarly, we can get the desired bound on $\|\hat{R}_2 - R_2\|_F^2$ over the event E_{n_1,n_2}. The proof is done.

References

[1] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 670–688. IEEE, 2015.

[2] Xing Chen, Ming-Xi Liu, and Gui-Ying Yan. Drug–target interaction prediction by random walk on the heterogeneous network. *Molecular BioSystems*, 8(7):1970–1978, 2012.

[3] Ying Ding. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of informetrics*, 5(1):187–203, 2011.

[4] Xiaowen Dong, Pascal Frossard, Pierre Vandergheynst, and Nikolai Nefedov. Clustering with multi-layer graphs: A spectral perspective. *IEEE Transactions on Signal Processing*, 60(11): 5820–5831, 2012.

[5] Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An l_∞ eigenvector perturbation bound and its application to robust covariance estimation. *arXiv preprint arXiv:1603.03516*, 2016.

[6] Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Community detection in degree-corrected block models. *arXiv preprint arXiv:1607.06993*, 2016.

[7] Qiuyi Han, Kevin S Xu, and Edoardo M Airoldi. Consistent estimation of dynamic and multi-layer networks. *arXiv preprint arXiv:1410.8597*, 2014.

[8] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[9] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[10] Rodolphe Jenatton, Nicolas L Roux, Antoine Bordes, and Guillaume R Obozinski. A latent factor model for highly multi-relational data. In *Advances in Neural Information Processing Systems*, pages 3167–3175, 2012.

[11] Pengsheng Ji, Jiashun Jin, et al. Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812, 2017.

[12] Jiashun Jin et al. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.

[13] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[14] Daniel B Larremore, Aaron Clauset, and Abigail Z Jacobs. Efficiently inferring community structure in bipartite networks. *Physical Review E*, 90(1):012805, 2014.

[15] Zhen Li, Zhisong Pan, Yanyan Zhang, Guopeng Li, and Guyu Hu. Efficient community detection in heterogeneous social networks. *Mathematical Problems in Engineering*, 2016, 2016.

[16] Zhenping Li, Rui-Sheng Wang, Shihua Zhang, and Xiang-Sun Zhang. Quantitative function and algorithm for community detection in bipartite networks. *Information Sciences*, 367: 874–889, 2016.

[17] Xin Liu, Weichu Liu, Tsuyoshi Murata, and Ken Wakita. A framework for community detection in heterogeneous multi-relational networks. *Advances in Complex Systems*, 17(06): 1450018, 2014.

[18] Travis Martin, Brian Ball, Brian Karrer, and MEJ Newman. Coauthorship and citation patterns in the physical review. *Physical Review E*, 88(1):012814, 2013.

[19] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *science*, 328 (5980):876–878, 2010.

[20] Subhadeep Paul and Yuguo Chen. Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics*, 10(2): 3807–3870, 2016.

[21] Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.

[22] Srijan Sengupta and Yuguo Chen. Spectral clustering in heterogeneous networks. *Statistica Sinica*, pages 1081–1106, 2015.

[23] Gilbert W Stewart. Matrix perturbation theory. 1990.

[24] Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD Explorations Newsletter*, 14(2):20–28, 2013.

[25] Yoshihiro Yamanishi, Masaaki Kotera, Minoru Kanehisa, and Susumu Goto. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, 26(12):i246–i254, 2010.

[26] Fan Yang, Jinbo Xu, and Jianyang Zeng. Drug-target interaction prediction by integrating chemical, genomic, functional and pharmacological data. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 148. NIH Public Access, 2014.

[27] Yunpeng Zhao, Elizaveta Levina, Ji Zhu, et al. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4): 2266–2292, 2012.