Statistica Sinica **25** (2015), 1081-1106
doi:http://dx.doi.org/10.5705/ss.2013.231

SPECTRAL CLUSTERING IN HETEROGENEOUS NETWORKS

Srijan Sengupta and Yuguo Chen

University of Illinois at Urbana-Champaign

Abstract: Many real-world systems consist of several *types* of entities, and heterogeneous networks are required to represent such systems. However, the current statistical toolbox for network data can only deal with homogeneous networks, where all nodes are supposed to be of the same type. This article introduces a statistical framework for community detection in heterogeneous networks. For modeling heterogeneous networks, we propose heterogeneous versions of both the classical stochastic blockmodel and the degree-corrected blockmodel. For community detection, we formulate heterogeneous versions of standard spectral clustering and regularized spectral clustering. We demonstrate the theoretical accuracy of the proposed heterogeneous methods for networks generated from the proposed heterogeneous models. Our simulations establish the superiority of proposed heterogeneous methods over existing homogeneous methods in finite networks generated from the models. An analysis of the DBLP four-area data demonstrates the improved accuracy of the heterogeneous method over the homogeneous method in identifying research areas for authors.

Key words and phrases: Clustering, community detection, degree-corrected blockmodel, heterogeneous network, network analysis, regularized spectral clustering, spectral clustering, stochastic blockmodel.

1. Introduction

Many complex systems in today's world consist, at an abstract level, of *agents* who *interact* with one another. This general agent-interaction framework describes many interesting and important systems, such as social interpersonal systems (Milgram (1967)), protein interaction systems (Gavin et al. (2002)), power grids (Watts and Strogatz (1998)), and the World Wide Web (Huberman and Adamic (1999)), to name a few. Networks provide a convenient and unified way of representing such systems. It is important to develop methodology for network data and, accordingly, the science of network data has received attention from scientists in various academic fields. A holistic introduction to the interdisciplinary study of networks can be found in Newman (2010). Statistically oriented overview of networks can be found in Goldenberg et al. (2010) and Kolaczyk (2009).

The early approach to network modeling, the random graph model of Erdös and Rényi (1959), assumed that all agents behave in identical fashion. The observed dissimilarity in agent behavior was assigned to random fluctuations. This explanation is not always appropriate, particularly when the network displays structured dissimilarities in agent behavior. At the other end of the modeling spectrum, one might wish to capture the observed variation in agent behavior by assigning a separate model to each individual agent. However, this is impractical for networks beyond a certain size, and also unnecessary.

Observed networks often exhibit a *patterned dissimilarity* that lies somewhere between completely identical agent behavior and completely unequal agent behavior. Agents are often found to cluster into groups or communities that display similar behavior, while agents from different communities behave differently. The identification of this network structure, called *community detection*, is an important problem. Community detection has important interpretation; communities often turn out to be groups of agents that share common properties and/or play similar roles within the network. For example, in Jonsson et al. (2006), the communities in a protein interaction network turned out to be functional groups (proteins having the same or similar function), with important implications for cancer research. Fortunato (2010) provides a multidisciplinary exposition on community detection in networks.

The currently available methodologies for network data usually consider *homogeneous* networks consisting of nodes that represent objects of the same type, and that all links in the network represent the same type of relation. For example, a friendship network like Facebook has nodes representing persons or users, and links representing friendship between users. However, many observed systems are *heterogeneous* in that there are different *types* of agents, and various kinds of interactions in the system. Typically, for each node or link it is known what the type is, and a heterogeneous network contains this type information. For example, in Facebook, nodes can represent various types of entities like users, events, groups, celebrity pages, photos, and so on. Accordingly, there can be various types of links: friendship link between two users, membership link between users and groups, fan (or *like*) link between users and celebrities, attendance link between users and events, *tag* link between a photo and an user, and so on. The homogeneous 'friendship network' representation, that was mentioned earlier, effectively represents only a sub-system of this system, consisting only of 'user' nodes and 'friendship' links.

To analyze a heterogeneous network using the current toolbox of homogeneous methods and homogeneous models, there are two options — either consider a homogeneous sub-network of the original network, or treat the heterogeneous network as a homogeneous network, suppressing the type information available in the data. In the first approach, there is loss of useful information. In the

second approach the results might be meaningless as nodes of different types are grouped into the same community, or the procedure might not work well due to the presence of different types of nodes.

For example, consider a heterogeneous Facebook network consisting of two types of nodes — users and events, and two kinds of links: user-user or *friendship* links, and user-event or *attendance* links. Suppose network data in this form is available for users and events corresponding to 10 universities, and the problem of interest is to assign users to their universities using a clustering procedure. Using the first option, one must carry out the analysis based on the user-user network only, dumping the event nodes and user-event links. In this context the dumped data can be quite important in predicting university affiliation. Using the second option, one treats the entire network as a homogeneous network and carries out a clustering of both users and events. However, users and events behave in very different ways, and the clustering algorithm might not work well since it is trying to cluster these different entities into the same clusters by comparing their behavior. Using K-means intuition, the 'distance' between an user and an event, both affiliated to the same university, might be too large.

Thus, community detection in heterogeneous network data cannot be satisfactorily carried out by applying homogeneous models and methodologies. A preferable approach is to have a procedure that uses the entire heterogeneous information, identifies the fact that users and events are different types of entities, and clusters nodes from different types *separately but simultaneously* into 10 user clusters and 10 event clusters. Since this procedure compares events to events and users to users, the clustering should work much better.

Heterogeneous networks have begun to receive attention from various scientific communities, particularly the computer science research community (Sun and Han (2012)). We provide a statistical framework to deal with heterogeneous network data, proposing heterogeneous versions of the classical stochastic blockmodel and the degree-corrected blockmodel of Karrer and Newman (2011). For community detection in heterogeneous networks, we formulate heterogeneous versions of standard spectral clustering and regularized spectral clustering. We demonstrate the theoretical accuracy of the proposed heterogeneous methods for networks generated from the proposed heterogeneous models in the asymptotic framework of Qin and Rohe (2013).

As an application of our methods, we analyze a large bibliographical network from DBLP with the objective of identifying research area of authors. A natural choice of network would be the homogeneous co-authorship network with authors as nodes, but we find that homogeneous clustering applied on the co-authorship network performs rather poorly, with an accuracy comparable to random assignment. Here, interpreting the bibliographical network as a heterogeneous network

(with authors, papers, and conferences treated as different types of nodes), the heterogeneous clustering method performs quite accurate community detection.

The rest of the article is organized as follows. Section 2 outlines basic graph theoretical notation that is used throughout the article. Section 3 reviews homogeneous blockmodels and introduces heterogeneous versions of them. Section 4 discusses standard and regularized spectral clustering algorithms and presents modified versions of these algorithms that are appropriate for heterogeneous networks. Section 5 provides a brief outline of the asymptotic framework of Qin and Rohe (2013), and demonstrates the asymptotic accuracy of the heterogeneous algorithms under heterogeneous models, using this framework. Section 6 presents simulation studies demonstrating various circumstances under which the heterogeneous methods can provide significant improvements in clustering accuracy over the homogeneous methods. Section 7 presents an example of the superiority of the heterogeneous method over the homogeneous method, using the DBLP four-area dataset. The article concludes with discussion in Section 8.

2. Graph Theoretic Notation

A network is represented as a *graph* $G = (V, E)$ consisting of *nodes* (or *vertices*) that comprise the set V, and *links* (or *edges*) that make up the set E. Every link has two endpoints in the set of nodes, and is said to *connect* or *join* the two nodes. The two endpoints of a link are also said to be *adjacent* to each other, or *neighbors*. An unweighted, undirected graph containing no self-loops or multiple edges is called a *simple graph*. The *degree* d_v of a node v in a graph G is the number of nodes adjacent to v. A *degree sequence* is a list of degrees of a graph in non-increasing order, say $d_1 \geq d_2 \geq \cdots \geq d_n$.

An *adjacency matrix* $\mathbf{A}$ for a graph with N nodes is an N-by-N matrix whose $(i, j)^{\text{th}}$ entry gives the number of links from the i^{th} node to the j^{th} node. We treat simple graphs only, and hence the adjacency matrix is symmetric, consists only of 0's and 1's, and all its diagonal entries are zero.

The *graph Laplacian* $\mathbf{L}$ is a matrix frequently used in network analysis. There are several ways of defining the Laplacian; in this article it is defined as

$$\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}, \tag{2.1}$$

where $\mathbf{A}$ is the adjacency matrix and $\mathbf{D}$ is the degree matrix, a diagonal matrix whose i^{th} diagonal element is the degree of node i. This version of the Laplacian is often referred to as the symmetric normalised Laplacian.

3. Stochastic Blockmodel and Degree Corrected Blockmodel for Heterogeneous Networks

Lorrain and White (1971) were the first to introduce blockmodels, in association with the deterministic concept of *structural equivalence*, where two nodes of

a network are considered equivalent if they have the same set of neighbors. Holland, Laskey, and Leinhardt (1983) and Fienberg, Meyer, and Wasserman (1985) generalized this equivalence concept to a probabilistic setting, calling it *stochastic equivalence*. In contrast to structural equivalence which is defined with respect to the observed network itself, stochastic equivalence is defined with respect to the conceptual model that generates the observed network.

Definition 1. Two nodes in a network are said to be *stochastically equivalent* if the probability of any event pertaining to the network remains unchanged by exchanging the node labels.

For a homogeneous network, two nodes (say, 1 and 2) are stochastically equivalent according to this definition if they have the same probability of being linked to *any* third node (say 3).

3.1. Homogeneous model

Consider a simple graph $G = (V, E)$ with N nodes, and let $\mathbf{A}$ be its adjacency matrix. Here $\mathbf{A}$ is a symmetric 0-1 matrix, and its diagonal entries are all zero. Under the K-block stochastic blockmodel, there are K blocks and each node belongs to one of these blocks. Let $\mathbf{M}$ denote the N-by-K block membership matrix with $\mathbf{M}(i,k) = 1$ if node i is in the k^{th} block, and $\mathbf{M}(i,k) = 0$ otherwise. Then for $i < j$, under the stochastic blockmodel (SBM), the $A(i,j)$ are Bernoulli random variables, with

$$E[\mathbf{A}(i,j) \mid \mathbf{M}] = \mathbf{M}(i,\cdot)\mathbf{P}\mathbf{M}(j,\cdot)', \tag{3.1}$$

where $\mathbf{P}$ is the K-by-K matrix of link probabilities: $\mathbf{P}(a,b)$ is the probability that a node in block a is linked to another node in block b. Edges are conditionally independent given the membership matrix $\mathbf{M}$.

Model (3.1) essentially means that if nodes i and i' come from the same block, then they are stochastically equivalent, since exchanging the node labels i and i' does not affect the probability of any event in the network.

Stochastic equivalence has nodes in the same block with identical degree distributions, and this can be an unrealistic assumption. The degree-corrected blockmodel (DCBM) proposed by Karrer and Newman (2011) adds degree scaling parameters θ_i for each node to allow for a broad degree distribution. Then for $i < j$, under the DCBM the $A(i,j)$ are Bernoulli random variables, with

$$E[\mathbf{A}(i,j) \mid \mathbf{M}] = \mathbf{\Theta}(i,\cdot)\mathbf{M}\mathbf{P}\mathbf{M}'\mathbf{\Theta}(\cdot,j), \tag{3.2}$$

where $\mathbf{\Theta}$ is an N-by-N diagonal matrix with $\mathbf{\Theta}(i,i) = \theta_i$, the degree parameter of the i^{th} node, and all other parameters have meaning as in (3.1). Note that the SBM is a special case of the DCBM when all nodes in the same block have equal value of the θ_i, the degree parameter.

3.2. Heterogeneous model

A heterogeneous network has nodes of different *types* with different roles in the network. Links in the network are also of different kinds, depending upon the types of the nodes they link. A blockmodel for heterogeneous networks should allow the link probabilities to change not only by block but also by node type. We propose a model for accommodating this.

Consider a K-block heterogeneous network with N nodes of T different types. We divide each block into T sub-blocks for different types of nodes such that each type-block combination is represented by a separate sub-block. Let $\mathbf{M}$ be the N-by-TK *sub-block* membership matrix, with $\mathbf{M}(i, t \times k) = 1$ if node i is of the t^{th} type and belongs to the k^{th} block, and $\mathbf{M}(i, t \times k) = 0$ otherwise, for $t = 1, \ldots, T$ and $k = 1, \ldots, K$. Let $\mathbf{P}$ be the TK-by-TK matrix of link probabilities. Then $\mathbf{P}$ has the following structure:

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \ldots & \mathbf{P}_{1T} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \ldots & \mathbf{P}_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{T1} & \mathbf{P}_{T2} & \ldots & \mathbf{P}_{TT} \end{pmatrix},$$

where $\mathbf{P}_{st}$ is the K-by-K matrix of probabilities for type s-type t links. Thus, $\mathbf{P}_{st}(a, b)$ represents the probability that a node of the s^{th} type and belonging to block a is linked to another node of the t^{th} type and belonging to block b.

For $i < j$, $\mathbf{A}(i, j)$ is a Bernoulli random variable, with

$$E[\mathbf{A}(i, j) \mid \mathbf{M}] = \mathbf{M}(i, \cdot)\mathbf{P}\mathbf{M}(j, \cdot)' \tag{3.3}$$

for the heterogeneous stochastic blockmodel (Het-SBM) and

$$E[\mathbf{A}(i, j) \mid \mathbf{M}] = \mathbf{\Theta}(i, \cdot)\mathbf{M}\mathbf{P}\mathbf{M}'\mathbf{\Theta}(\cdot, j) \tag{3.4}$$

for the heterogeneous degree-corrected blockmodel (Het-DCBM).

This complicated representation of link probabilities is necessary, because link probabilities vary not only by block, but also by type; $\mathbf{P}_{st}(a, b)$ vary not only with a and b but also with s and t.

For illustration, consider a toy example for Het-SBM with number of types $T = 2$, the number of blocks $K = 3$ and the number of nodes $N = 30$, with 5 type 1 nodes and 5 type 2 nodes in each block, and link probability matrix

$$\mathbf{P} = \begin{pmatrix} 0.75 & 0.25 & 0.25 & 0.90 & 0.00 & 0.00 \\ 0.25 & 0.75 & 0.25 & 0.00 & 0.90 & 0.00 \\ 0.25 & 0.25 & 0.75 & 0.00 & 0.00 & 0.90 \\ & & & & & \\ 0.90 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.90 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.90 & 0.00 & 0.00 & 0.00 \end{pmatrix}.$$

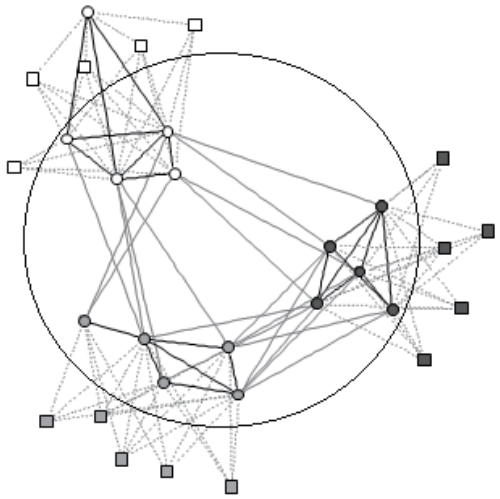


Figure 1. Sample heterogeneous network with $T = 2, K = 3$ and $N = 30$, with 5 type 1 nodes (circles) and 5 type 2 nodes (squares) in each block. Solid lines (black for intra-block, gray for inter-block) represent type 1-type 1 links, while dotted gray lines represent type 1-type 2 links. The homogeneous type 1-type 1 subnetwork is approximately enclosed in the circle.

Link probabilities vary prominently across blocks as well as types in this model. Type 1 nodes are strongly homophilic (intra-community links are much more likely than inter-community links), while type 2 nodes are not linked. The type 1-type 2 links have even stronger homophily; type 1 and type 2 nodes belonging to the same block are very likely to be connected, while inter-block, inter-type links are not present. This model is an exaggerated representation of the user-event heterogeneous Facebook system mentioned in the introduction. In Figure 1, type 1 nodes and type 2 nodes are clearly different in their roles in the network, nevertheless they form close-knit communities. Visually, it appears that community structure is stronger in the entire network, compared to the homogeneous type 1-type 1 subnetwork. This toy example gives a visual intuition of how community discovery might be more accurate in the presence of heterogeneous information.

4. Spectral Clustering and Regularized Spectral Clustering

4.1. Homogeneous clustering

Consider a homogeneous network with N nodes and let $\mathbf{A}$ be its adjacency matrix. Assuming a correctly specified K-block blockmodel structure for this

network, the standard spectral clustering algorithm assigns the N nodes to K clusters in the following steps.

Homogeneous Spectral Clustering Algorithm (Hom-SC)

1. Given the adjacency matrix $\mathbf{A}$, calculate the graph Laplacian $\mathbf{L}$ by (2.1).
2. Find orthonormal eigenvectors $\mathbf{X}_1, \ldots, \mathbf{X}_K$ corresponding to the K eigenvalues of $\mathbf{L}$ that are largest in absolute value. Put them into the N-by-K matrix $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_K]$.
3. Carry out a K-means clustering with the N rows of matrix $\mathbf{X}$, creating a K-partition of the index set $\{1, \ldots, N\}$.
4. For each i, assign the i^{th} node to the k^{th} cluster if the i^{th} row was assigned to the k^{th} cluster in Step 3.

Recent work by Amini et al. (2013) and Jin (2012) demonstrate that homogeneous spectral clustering does not work very well in sparse homogeneous networks with wide degree distribution. Chaudhuri, Chung, and Tsiatas (2012) proposed a regularized version of the graph Laplacian for sparse networks and it was shown by Qin and Rohe (2013) that a normalized variant of spectral clustering on this regularized Laplacian has superior theoretical properties under the degree corrected stochastic blockmodel. In this context we mention Joseph and Yu (2013) for their in-depth analysis of the performance of a slightly different version of regularized spectral clustering.

For a regularizer $\tau \geq 0$, take the regularized degree matrix $\mathbf{D}_\tau = \mathbf{D} + \tau \mathbf{I}$ and the regularized graph Laplacian

$$\mathbf{L}_\tau = \mathbf{D}_\tau^{-1/2} \mathbf{A} \mathbf{D}_\tau^{-1/2}. \tag{4.1}$$

Following Qin and Rohe (2013), we set τ equal to the average node degree of the network in all applications of regularized spectral clustering (homogeneous and heterogeneous), for simulations as well as data analysis. The regularized spectral clustering algorithm assigns the N nodes to K clusters in the following steps.

Homogeneous Regularized Spectral Clustering Algorithm (Hom-RSC)

1. Given the adjacency matrix $\mathbf{A}$ and regularizer $\tau \geq 0$, calculate regularized graph Laplacian $\mathbf{L}_\tau$ by (4.1).
2. Find orthonormal eigenvectors $\mathbf{X}_1, \ldots, \mathbf{X}_K$ corresponding to the K eigenvalues of $\mathbf{L}_\tau$ that are largest in absolute value. Put them into the N-by-K matrix $\mathbf{X}_\tau = [\mathbf{X}_1, \ldots, \mathbf{X}_K]$.
3. Normalize each row of $\mathbf{X}_\tau$ to have unit norm, forming the N-by-K matrix $\mathbf{X}_\tau^*$ given by $\mathbf{X}_\tau^*(i,j) = \mathbf{X}_\tau(i,j) \Big/ \sqrt{\sum_j \mathbf{X}_\tau(i,j)^2}$.

4. Carry out a K-means clustering with the rows of matrix $\mathbf{X}_\tau^*$, creating a K-partition of the index set $\{1, \ldots, N\}$.
5. For each i, assign the i^{th} node to the k^{th} cluster if the i^{th} row was assigned to the k^{th} cluster in Step 4.

4.2. Heterogeneous clustering

For a T-type heterogeneous network, there are TK clusters (since each type-block combination represents a cluster), but for each node, the type information is already known. So essentially there are T cluster assignment problems — to assign the n_1 type 1 nodes into K clusters, the n_2 type 2 nodes into K separate clusters, and so on. This can be achieved by carrying out T *simultaneous but separate* K-means clustering procedures. We now present heterogeneous versions of the Hom-SC and Hom-RSC algorithms based on this idea.

Heterogeneous Spectral Clustering Algorithm (Het-SC)

1. Given the adjacency matrix $\mathbf{A}$, calculate the graph Laplacian $\mathbf{L}$ by (2.1).
2. Find orthogonal eigenvectors $\mathbf{X}_1, \ldots, \mathbf{X}_{TK}$ corresponding to the TK eigenvalues of $\mathbf{L}$ that are largest in absolute value. Put them into the N-by-TK matrix $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_{TK}]$.
3. For each $t = 1, \ldots, T$, select the n_t rows of $\mathbf{X}$ that correspond to nodes of type t, and carry out separate K-means clustering for each selection, creating a TK-partition of the index set $\{1, \ldots, N\}$.
4. For each i, assign the i^{th} node to the r^{th} cluster if the i^{th} row was assigned to the r^{th} cluster in Step 3.

Heterogeneous Regularized Spectral Clustering Algorithm (Het-RSC)

1. Given the adjacency matrix $\mathbf{A}$ and regularizer $\tau \geq 0$, calculate the regularized graph Laplacian $\mathbf{L}_\tau$ by (4.1).
2. Find orthogonal eigenvectors $\mathbf{X}_1, \ldots, \mathbf{X}_{TK}$ corresponding to the TK eigenvalues of $\mathbf{L}_\tau$ that are largest in absolute value. Put them into the N-by-TK matrix $\mathbf{X}_\tau = [\mathbf{X}_1, \ldots, \mathbf{X}_{TK}]$.
3. Normalize each row of $\mathbf{X}_\tau$ to have unit norm, forming the N-by-TK matrix $\mathbf{X}_\tau^*$ given by $\mathbf{X}_\tau^*(i,j) = \mathbf{X}_\tau(i,j) \Big/ \sqrt{\sum_j \mathbf{X}_\tau(i,j)^2}$.
4. For each $t = 1, \ldots, T$, select the n_t rows of $\mathbf{X}_\tau^*$ that correspond to nodes of type t, and carry out separate K-means clustering for each selection, creating a TK-partition of the index set $\{1, \ldots, N\}$.
5. For each i, assign the i^{th} node to the r^{th} cluster if the i^{th} row was assigned to the r^{th} cluster in Step 4.

In the next section we justify the use of the Het-RSC algorithm under the Het-DCBM model and the Het-SC algorithm under the Het-SBM model.

5. Convergence of Heterogenous Spectral Clustering

We outline an asymptotic theory for the convergence of the Hom-RSC algorithm under the Hom-DCBM, and propose a similar result for the Het-RSC algorithm under the Het-DCBM. Convergence of the Het-SC algorithm under the Het-SBM follows as a special case. See Qin and Rohe (2013) for technical details for the homogeneous case.

For the Hom-DCBM (3.2), define $\mathcal{A} = \mathbf{\Theta M P M' \Theta}$ and let $\mathcal{D}$ be the diagonal matrix of expected degrees, $\mathcal{D}(i,i) = \sum_j \mathcal{A}(i,j)$. Set $\mathcal{D}_\tau = \mathcal{D} + \tau \mathbf{I}$ and let $\mathcal{L}_\tau = \mathcal{D}_\tau^{-1/2} \mathcal{A} \mathcal{D}_\tau^{-1/2}$ be the population version of the regularized graph Laplacian (4.1). Under a K-block Hom-DCBM, $\mathcal{L}_\tau$ has exactly K non-zero eigenvalues. Let $\mathcal{X}_\tau$ be the N-by-K matrix of the corresponding eigenvectors, with $\mathcal{X}_\tau^*$ the row-normalized version of $\mathcal{X}_\tau$. The main idea is to interpret the clustering algorithm as an estimation procedure for $\mathbf{X}^*$ with $\mathcal{X}_\tau^*$ as the parameter.

Let $\delta = \min_{i=1,\dots,N} \mathcal{D}(i,i)$ be the minimum expected degree, and let λ be the magnitude of the smallest non-zero eigenvalue of $\mathcal{L}_\tau$ in magnitude. Let $\gamma = \min_{i=1,\dots,N}\{\min\{||\mathbf{X}_\tau(i,\cdot)||_2, ||\mathcal{X}_\tau(i,\cdot)||_2\}$ be the length of the shortest row in $\mathcal{X}_\tau$ and $\mathbf{X}_\tau$, where $||x||_2$ represents the L^2 norm of the vector x. Assume that for some $\epsilon > 0$ and sufficiently large N,

$$\text{(A1) } \delta + \tau > 3\log(4N/\epsilon) \quad \text{ and } \quad \text{(A2) } \lambda \geq 8\sqrt{\frac{3K\log(4N/\epsilon)}{\delta+\tau}}.$$

Theorem 4.2 of Qin and Rohe (2013) states: when (A1) and (A2) hold, then

$$||\mathbf{X}_\tau - \mathcal{X}_\tau \mathbf{O}||_F \leq c_0 \frac{1}{\lambda}\sqrt{\frac{K\log(4N/\epsilon)}{\delta+\tau}}, \tag{5.1}$$

$$||\mathbf{X}_\tau^* - \mathcal{X}_\tau^* \mathbf{O}||_F \leq c_0 \frac{1}{\gamma\lambda}\sqrt{\frac{K\log(4N/\epsilon)}{\delta+\tau}}, \tag{5.2}$$

for some constant c_0 with probability at least $1-\epsilon$, where $\mathbf{O}$ is an orthonormal rotation, and $||\cdot||_F$ is the Frobenius norm of a matrix, $||\mathbf{B}||_F = \sqrt{\sum_i \sum_j |\mathbf{B}(i,j)|^2}$.

The next step is to translate this accuracy in the estimation of $\mathcal{X}_\tau^*$ into accurate clustering of nodes. Lemma 3.3 of Qin and Rohe (2013) shows that $\mathcal{X}_\tau^*$ can be written as $\mathcal{X}_\tau^* = \mathbf{MB}$, where $\mathbf{B}$ is a K-by-K non-singular matrix. The membership matrix $\mathbf{M}$ has exactly K unique rows, so $\mathcal{X}_\tau^*$ also has exactly K unique rows. This implies that a K-means clustering applied to the rows of $\mathcal{X}_\tau^*$ would perfectly identify the block membership of all nodes in the network. Given the asymptotic closeness between $\mathbf{X}_\tau^*$ and $\mathcal{X}_\tau^*$ from (5.2), as the clustering output

from $\mathcal{X}_\tau^*$ is perfect, the clustering output from $\mathbf{X}_\tau^*$ is expected to approach that perfect accuracy in an asymptotic sense.

To formalize this intuition, a tractable definition of misclustering is required. In Step 4 of the regularized spectral clustering algorithm, the N rows of $\mathbf{X}_\tau^*$ are subjected to a K-means clustering that assigns each row to a cluster, and each cluster thus formed has a centroid. Let $\mathbf{C}$ be the N-by-K matrix with $\mathbf{C}(i,\cdot)$ the centroid corresponding to the i^{th} row of $\mathbf{X}_\tau^*$. Then $\mathcal{X}_\tau^*(i,\cdot)$ is the parameter centroid corresponding to the i^{th} node, while the estimated centroid is $\mathbf{C}(i,\cdot)$. It is therefore reasonable to consider the i^{th} node to be correctly clustered if the estimated centroid is closer to the correct parameter centroid than the remaining $K-1$ incorrect parameter centroids, and it is misclustered if the estimated centroid is closer to some incorrect parameter centroid than the correct parameter centroid.

Definition 2. The set of misclustered nodes $\mathcal{E}$ is defined as

$$\mathcal{E} = \{i : \exists\ j \neq i\ s.t.\ ||\mathbf{C}(i,\cdot) - \mathcal{X}_\tau^*(i,\cdot)\mathbf{O}||_2 > ||\mathbf{C}(i,\cdot) - \mathcal{X}_\tau^*(j,\cdot)\mathbf{O}||_2\}\,. \tag{5.3}$$

For some $\epsilon > 0$ and sufficiently large N, suppose (A1) and (A2) hold. Then Theorem 4.4 of Qin and Rohe (2013) states that, with probability at least $1-\epsilon$,

$$|\mathcal{E}| \leq c_1 \frac{K \log(N/\epsilon)}{\gamma^2 \lambda^2 (\delta+\tau)} \tag{5.4}$$

for some constant c_1.

We extend these ideas to the Het-DCBM and the Het-RSC algorithm. For the T-type, K-block Het-DCBM from Section 3.2, $\mathbf{M}$ has TK unique rows, and $\mathbf{P}$ is TK-by-TK, since link probabilities are allowed to vary for each type-block combination. This model is structurally equivalent to a Hom-DCBM (3.2) with TK blocks. The interpretation of sub-blocks in a T-type, K-block heterogeneous model is different from that of blocks in a TK-block homogeneous model, but both models have the same mathematical structure. Consequently, the convergence result for row-normalized eigenvectors in (5.2) can be directly applied to the heterogeneous model.

The translation of estimation accuracy to clustering accuracy, however, does not extend directly from the homogeneous version to the heterogeneous version. The upper bound in (5.4) for the homogeneous case is derived from the fact that the matrix of cluster centroids, $\mathbf{C}$, is the minimizer of the K-means objective function $||\mathbf{X}_\tau^* - \mathbf{Y}||_F$, minimization being performed over the set of all N-by-K matrices $\mathbf{Y}$ having exactly K unique rows. The Het-RSC algorithm in Section 4.2 runs T separate K-means procedures on the T node types, thereby using a different objective function. However, after considering the modified objective function being minimized in Step 4 of the Het-RSC algorithm, we are able to prove a heterogeneous version of (5.4).

Theorem 1. *Consider a T-type, K-block Het-DCBM with n_t nodes of type t, and $N = \sum_{t=1}^{T} n_t$. For nodes of type t, define the set of misclustered nodes $\mathcal{E}_t$ as*

$$\mathcal{E}_t = \{i \in \textit{type } t : \exists\ j \in \textit{type } t\ \&\ j \neq i\ \textit{s.t.}\ ||\mathbf{C}(i,\cdot) - \mathcal{X}_\tau^*(i,\cdot)\mathbf{O}||_2 > ||\mathbf{C}(i,\cdot) - \mathcal{X}_\tau^*(j,\cdot)\mathbf{O}||_2\}, \quad (5.5)$$

and let $\gamma_t = \min_{i \in \textit{type } t}\{\min\{||\mathbf{X}_\tau(i,\cdot)||_2, ||\mathcal{X}_\tau(i,\cdot)||_2\}$ be the length of the shortest row of type t in $\mathcal{X}_\tau$ and $\mathbf{X}_\tau$, with λ, δ, τ defined as before. For some $\epsilon > 0$ and sufficiently large N, suppose (A1) and (A2) hold. Then with probability at least $1 - \epsilon$, for some constant c_1,

$$|\mathcal{E}_t| \leq c_1 \frac{K \log(N/\epsilon)}{\gamma_t^2 \lambda^2 (\delta + \tau)} \quad \textit{for } t = 1, \ldots, T. \quad (5.6)$$

The proof of Theorem 1 is in the Appendix.

Remark 1. Assumption (A1) requires a lower bound on the smallest regularized expected degree $\delta + \tau$. This emphasizes the importance of regularization, as it allows expected node degrees to be low, as long as they are complemented by the regularizer. Assumption (A2) requires that the smallest non-zero eigenvalue of $\mathcal{L}_\tau$ in magnitude does not decay to zero too fast. The number of types, T, is arbitrary but fixed. The number of blocks, K, is allowed to increase with n_t and N as long as (A1) and (A2) hold true — thus its allowable rate of increase depends on the large sample behavior of the quantities δ, τ, and λ.

Theorem 1 provides a separate bound for each node type. Under (A1) and (A2), for a given type t and for sufficiently large N, the quantity on the right side of (5.6) is $O(1/\gamma_t^2)$. Therefore as $n_t \to \infty$ the asymptotic bound on the number of misclustered nodes depends on γ_t. When γ_t decays at a rate slower than $\sqrt{1/n_t}$, the error rate $|\mathcal{E}_t|/n_t$ goes to zero. The bound deteriorates when γ_t decays to zero faster than $\sqrt{1/n_t}$. It is plausible that the bound goes to zero for certain node types, but not for others depending on the behavior of γ_t for different node types.

Remark 2 (Application to Het-SC under Het-SBM)**.** The convergence of Hom-SC under Hom-SBM was first established by Rohe, Chatterjee, and Yu (2011) under the assumption of a dense network model. Their results can be extended from the homogeneous setting to the heterogeneous setting, but would restrict the application to the dense network case. The framework of Qin and Rohe (2013) allows for sparse networks in a broader class of degree-corrected models, and the results can be readily applied to Het-SC under Het-SBM, as outlined below.

The main difference between the Het-SC algorithm and the Het-RSC algorithm is that the former does not have regularization or normalization steps. The Het-SBM is a special case of the Het-DCBM, when the degree parameters θ_i are equal. In this special case, the model eigenvector matrix $\mathcal{X}_\tau$ already has exactly TK distinct rows (applying Lemma 3.3 of Qin and Rohe (2013)) corresponding to the TK sub-blocks. Hence, row-normalization is not required — we can cluster the rows of $\mathbf{X}_\tau$ directly and use the result in (5.1). Further, we can proceed with no regularization. In doing so the advantage of regularization is lost, and the model is required to satisfy restricted version of (A1) and (A2):

$$\text{(A1}')\ \delta > 3\log\left(\frac{4N}{\epsilon}\right) \quad \text{and} \quad \text{(A2}')\ \lambda \geq 8\sqrt{\frac{3K\log(4N/\epsilon)}{\delta}}.$$

Here λ is the magnitude of the smallest non-zero eigenvalue of $\mathcal{L}$ (the unregularized Laplacian) in magnitude.

Theorem 2. *Consider a T-type, K-block Het-SBM with n_t nodes of type t, and $N = \sum_{t=1}^{T} n_t$. Let $\mathbf{C}$ denote the matrix of cluster centroids resulting from Het-SC, and the set of misclustered nodes of type t be defined as*

$$\mathcal{E}_t = \{i \in \text{type } t : \exists\ j \in \text{type } t\ \&\ j \neq i\ s.t. \\ ||\mathbf{C}(i,\cdot) - \mathcal{X}_{\tau=0}(i,\cdot)\mathbf{O}||_2 > ||\mathbf{C}(i,\cdot) - \mathcal{X}_{\tau=0}(j,\cdot)\mathbf{O}||_2\}. \quad (5.7)$$

Let λ and δ be defined as before, and $\tau = 0$. For some $\epsilon > 0$ and sufficiently large N, suppose (A1′) *and* (A2′) *hold. Then with probability at least $1-\epsilon$, for some constant c_1,*

$$|\mathcal{E}_t| \leq c_1 \frac{K\log(N/\epsilon)}{\lambda^2(\delta+\tau)} \quad \text{for } t = 1,\ldots,T. \quad (5.8)$$

The proof for Theorem 2 is essentially similar to that for Theorem 1, the only difference being the use of the eigenvector matrix $\mathbf{X}_{\tau=0}$ instead of its normalized version $\mathbf{X}^*_{\tau=0}$, and hence we skip the proof.

6. Simulation Results

We report on three simulation studies comparing the finite-sample performance of the homogeneous clustering algorithms with their heterogeneous counterparts in bi-type heterogeneous networks, and we studied both Het-SBM and Het-DCBM. Although Het-SBM is a special case of Het-DCBM, it is an important special case from a methodological perspective. Regularized spectral clustering adds two extra steps to standard spectral clustering — regularization (Step 1) and row-normalization (Step 3). The former aims to deal with sparsity, while the latter aims to deal with non-uniformity in expected node degrees. In

simulations, these features stem from degree parameters in Het-DCBM. For Het-SBM, the uniformity of expected node degrees makes both regularization and normalization unnecessary. Therefore, we studied the performance of Het-SC vs Hom-SC under networks generated from Het-SBM, and that of Het-RSC vs Hom-RSC in networks generated from Het-DCBM.

The class of Het-SBM models used for these simulations was $\mathcal{B}(K; s_1, s_2, p_1, r_1, p_2, r_2, p_3, r_3)$ where K is the number of blocks, and s_1 and s_2 are the number of type 1 and type 2 nodes per block, respectively. The probability matrix was given by $\mathbf{P} = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{pmatrix}$, where

$$\begin{aligned} \mathbf{P}_{11} &= p_1 \mathbf{1}_K \mathbf{1}_K' + r_1 \mathbf{I}_K, \\ \mathbf{P}_{22} &= p_2 \mathbf{1}_K \mathbf{1}_K' + r_2 \mathbf{I}_K, \\ \mathbf{P}_{12} \;=\; \mathbf{P}_{21} &= p_3 \mathbf{1}_K \mathbf{1}_K' + r_3 \mathbf{I}_K. \end{aligned}$$

Here $\mathbf{1}_K$ is a K-vector of 1's, and $\mathbf{I}_K$ is the K-by-K identity matrix. Thus, in the type 1-type 1 (type 2-type 2) homogeneous network, $p_1 (p_2)$ represents the inter-block link probability while $p_1 + r_1$ $(p_2 + r_2)$ is the intra-block link probability. The strength of homophily in the homogeneous networks is therefore determined by r_1 and r_2. For type 1-type 2 links, p_3 represents the inter-block, inter-type link probability and r_3 represents the strength of inter-type homophily. For Het-DCBM, we used the same values of $\mathbf{P}, K, s_1$, and s_2. The degree parameters θ_i were generated from the power law distribution

$$f(x) = \frac{\beta - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\beta}$$

with $x_{\min} = 1$ and shape parameter $\beta = 3$, and then scaled down so that the average $\bar{\theta}$ was 1. For a given parameter combination, the Het-DCBM can therefore be interpreted as a 'noisy' version of the Het-SBM, or conversely the Het-SBM can be interpreted as an 'averaged' version of the Het-DCBM. For illustration, the model used in the example in Section 3.2 had $K = 3$, $s_1 = s_2 = 5$, $p_1 = 0.25$, $r_1 = 0.50$, $p_2 = r_2 = p_3 = 0$, and $r_3 = 0.90$.

Our main objective in these simulations was to study how the improved accuracy of heterogeneous clustering over homogeneous clustering depends on r_3 for a fixed value of p_3. Ceteris paribus, higher values of r_3 make the type 1-type 2 links more strongly homophilic, and therefore make heterogeneous community detection easier. Hence, we expect Het-SC and Het-RSC to be increasingly accurate with increasing r_3, while Hom-SC and Hom-RSC are not affected by r_3. However, the actual improvement of heterogeneous clustering over homogeneous clustering depends on other parameters as well, particularly p_1, r_1, p_2, and r_2, which determine the strength of homophily for the homogeneous networks.

To capture these dynamics, we fixed $K = 3$, $s_1 = 100$, $s_2 = 50$, and $p_3 = 0.25$. Thus, we studied networks having a total of $N = 450$ nodes, of which 300 were of type 1 and 150 of type 2. The parameters p_1, p_2, r_1, and r_2 were set to various combinations, and for each combination, r_3 was increased from 0.10 to 0.50 in increments of 0.05. For each combination of parameters, error rates were estimated by averaging across 100 networks from Het-SBM and Het-DCBM.

How clustering performance was measured is important. Definitions (5.3), (5.5), and (5.7) introduced model-based quantification of misclustered nodes for the purpose of mathematical tractability, but these definitions require complete knowledge of the underlying model generating the network. For calculation of misclustering error in a real network, the true membership (ground truth) might be known (providing information about $\mathbf{M}$), but the other model parameters are unknown, and hence these definitions can not be evaluated for real networks. Accordingly, instead of this model-based quantity we used the data-based error rate: the proportion of nodes that got assigned to wrong clusters.

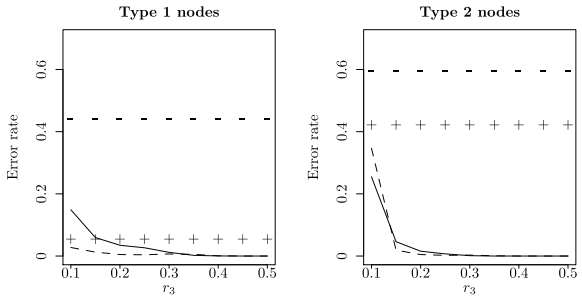
For a K-block network model, the *true* membership ($\mathbf{M}$) provides a K-partition, say $\mathfrak{P}$, of the nodes. Suppose there are two competing clustering algorithms that provide K-partitions, say $\mathfrak{P}_1$ and $\mathfrak{P}_2$. For each algorithm, consider the K-by-K overlap table $\mathbf{T}_i$ such that $\mathbf{T}_i(k, l)$ is the number of nodes that have been assigned to the k^{th} block according to $\mathfrak{P}$ and to the l^{th} block according to $\mathfrak{P}_i$. Then $\sum_k \mathbf{T}_i(k, k)$ is the total number of correctly clustered nodes according to $\mathfrak{P}_i$, but this ignores the identifiability of cluster labels $\mathfrak{P}_1$ and $\mathfrak{P}_2$. This issue can be resolved by permuting the columns of $\mathbf{T}_i$ to maximize $\sum_k \mathbf{T}_i(k, k)$.

Our simulations were then designed not to study the finite-sample behavior of the quantities involved in the asymptotic theory, but to study the relative accuracy of homogeneous and heterogeneous clustering methods in finite networks generated from Het-SBM and Het-DCBM.

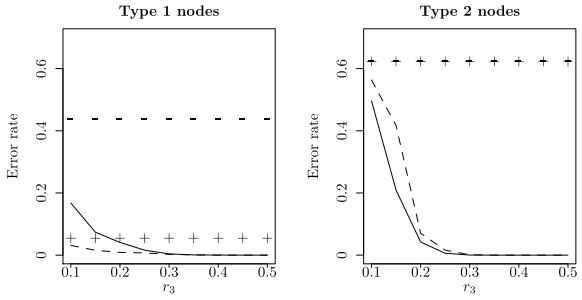
6.1. Simulation 1

In this simulation we studied homophilic networks where both types have similar inter-block and intra-block link probability. We used $p_1 = p_2 = 0.25$, and a single parameter, $r_1 = r_2 = r$, say, to govern the strength of homophily in the homogeneous networks. Values of $r = 0.10, 0.15$ were used to construct homogeneous networks of different strengths. The error rates from Het-SBM and Het-DCBM are plotted in Figure 2(a) and Figure 3(a), respectively.

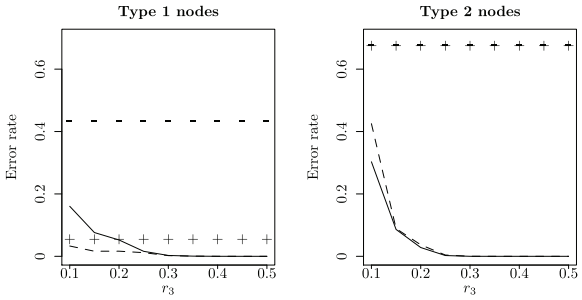
In both models homogeneous error rates for type 1 nodes were substantially lower than that for type 2, implying the effect of sample size or block size on error rates. Homogeneous error rates also decreased quite remarkably for both types as r was increased from 0.10 to 0.15, thereby increasing homophily between nodes. This phenomenon was more prominent in Het-SBM (Figure 2(a))



(a) Simulation 1: Both type 1-type 1 and type 2-type 2 networks are homophilic.

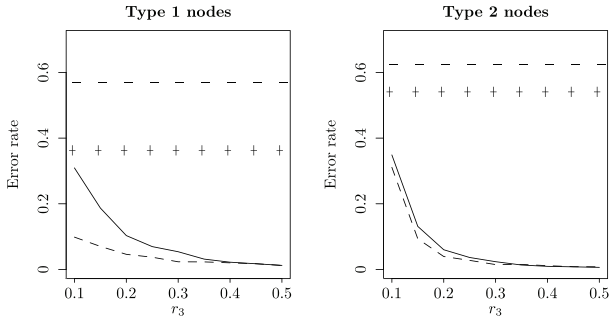


(b) Simulation 2: Type 1-type 1 networks have homophilic communities but type 2-type 2 networks do not have homophilic communities.

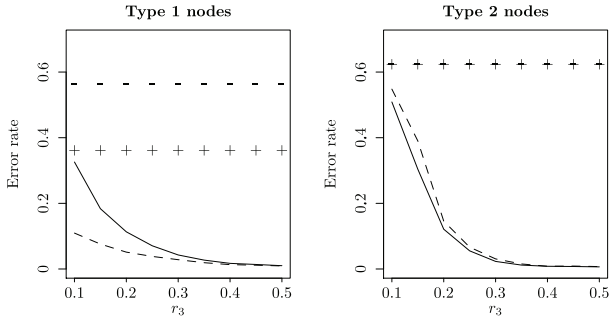


(c) Simulation 3: Homophilic type 1-type 1 networks and no type 2-type 2 links.

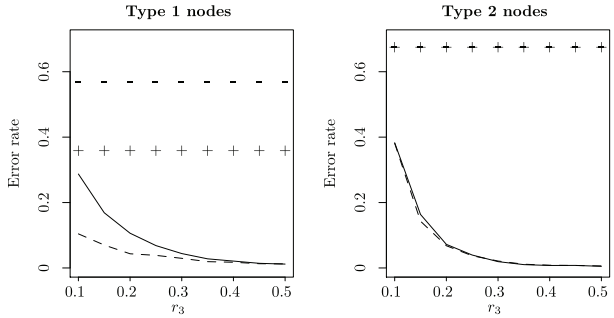
Figure 2. Comparison of homogeneous and heterogeneous clustering in Het-SBM from three simulation studies. For simulation 1, Hom-SC errors are represented as '-' for $r_1 = r_2 = 0.1$ and '+' for $r_1 = r_2 = 0.15$, while Het-SC errors are represented by solid lines for $r_1 = r_2 = 0.1$ and dashed lines for $r_1 = r_2 = 0.15$. For simulations 2 and 3, $r_2 = 0$, and Hom-SC errors are represented as '-' for $r_1 = 0.1$ and '+' for $r_1 = 0.15$, while Het-SC errors are represented by solid lines for $r_1 = 0.1$, and dashed lines for $r_1 = 0.15$. Note that type 2 nodes have a single Hom-SC error rate in both simulation 2 and simulation 3, as this error is not affected by r_1.



(a) Simulation 1: Both type 1-type 1 and type 2-type 2 networks are homophilic.



(b) Simulation 2: Type 1-type 1 networks have homophilic communities but type 2-type 2 networks do not have homophilic communities.



(c) Simulation 3: Homophilic type 1-type 1 networks and no type 2-type 2 links.

Figure 3. Comparison of homogeneous and heterogeneous clustering in Het-DCBM from three simulation studies. For simulation 1, Hom-RSC errors are represented as '-' for $r_1 = r_2 = 0.1$ and '+' for $r_1 = r_2 = 0.15$, while Het-RSC errors are represented by solid lines for $r_1 = r_2 = 0.1$ and dashed lines for $r_1 = r_2 = 0.15$. For simulations 2 and 3, $r_2 = 0$, and Hom-RSC errors are represented as '-' for $r_1 = 0.1$ and '+' for $r_1 = 0.15$, while Het-RSC errors are represented by solid lines for $r_1 = 0.1$, and dashed lines for $r_1 = 0.15$. Note that type 2 nodes have a single Hom-RSC error rate in both simulation 2 and simulation 3, as this error is not affected by r_1.

than Het-DCBM (Figure 3(a)). The most striking observation from Figures 2(a) and 3(a) was the improved accuracy of heterogeneous clustering over homogeneous clustering, for both types and both models. This comparative advantage increased with increasing r_3, but it was significant even for smaller values of r_3.

6.2. Simulation 2

A plausible scenario in heterogeneous networks is that the type 1-type 1 homogeneous network is homophilic but the type 2-type 2 network does not have homophilic community structure. To model this, we used $p_1 = p_2 = 0.25$ as before, but set the type 2 homophily parameter $r_2 = 0$ while the type 1 homophily parameter r_1 was increased from 0.1 to 0.15.

Figures 2(b) and 3(b) show that the heterogeneous methods are much more accurate for both node types. The improved accuracy over the homogeneous method is particularly remarkable for type 2 nodes, as the absence of homophily makes it difficult to assign communities to nodes on the basis of homogeneous information only. For example, consider a high school social network where students (type 1) form homophilic communities based on grades, but teachers (type 2) do not show homophily, rather they interact uniformly with other teachers. The heterogeneous student-teacher interaction is expected to be homophilic, as a student from a particular grade has more interaction with a teacher from the same grade, compared to a teacher from a different grade. In such a scenario, using a heterogeneous student-teacher network most likely performs better community detection for both students and teachers, compared to clustering the homogeneous student-student network or the homogeneous teacher-teacher network, even though teachers do not interact in homophilic fashion.

6.3. Simulation 3

Another plausible situation is that type 1-type 1 interactions are homophilic but there is no type 2-type 2 interaction at all. We used $p_1 = 0.25$ and increased r_1 from 0.1 to 0.15 as before, but set $p_2 = r_2 = 0$ so that there are no links between type 2 nodes. A motivation for this situation is the notional Facebook user-event heterogeneous network described in the introduction. While users (type 1) form a homophilic friendship network with universities as communities, there is no natural interaction between two events (type 2), implying a blank type 2-type 2 network. However, there is expected to be strong homophily in user-event interactions, and hence it is quite likely that the heterogeneous method will deliver a superior performance than the homogeneous method.

Figures 2(c) and 3(c) show that the heterogeneous method is indeed significantly superior to the homogeneous method, for both types. In this case, it is theoretically impossible to implement homogeneous spectral clustering on the

type 2-type 2 network, as the Laplacian for this network is a zero matrix, while the heterogeneous method delivered quite accurate clustering for type 2 nodes. For the sake of comparison, we used a flat homogeneous error rate of 2/3 (random allocation with $K = 3$ clusters) for type 2 nodes.

7. DBLP Four-Area Dataset Example

DBLP (Digital Bibliography & Library Project) is the authoritative computer science bibliography website, listing over two million articles. Gao et al. (2009) and Ji et al. (2010) extracted a connected subset of the DBLP data, containing bibliographical records from four research areas related to data mining: *database*, *data mining*, *information retrieval*, and *artificial intelligence*. The clustering problem of interest is to identify research area for authors. The original four-area dataset consists of 14,376 papers written by 14,475 authors, and presented at 20 conferences. The *ground truth* (true research area) is available for 4,057 authors, who account for 14,328 of these papers, covering all 20 conferences. Since error rates can be calculated only for labeled authors, our data analysis is based on this labeled subset of the data.

In the simulation studies of Section 6, we implemented Het-SC and Hom-SC on Het-SBM, and Het-RSC and Hom-RSC on Het-DCBM, backed by theoretical justification. However, in applications, we have to choose between standard and regularized spectral clustering, for both homogeneous and heterogeneous networks, on the basis of empirical features. In general, we expect regularization to work better if the network is sparse. Two distinguishing properties that are found in many large sparse networks (Girvan and Newman (2002)) are (i) a large number of nodes with low degrees, and (ii) power law behavior of degrees. We plot the histogram and log empirical tail distribution $\log_{10}(1 - \hat{F}(x))$ of node degrees in Figure 4 to investigate these properties. A heavily right-skewed histogram indicates property (i) and a roughly linear plot of $\log_{10}(1-\hat{F}(x))$ indicates property (ii). Accordingly, in the following analysis we chose regularization if the plots indicated sparsity.

7.1. Homogeneous author collaboration network

For homogeneous clustering, the natural network is the co-authorship network, where authors are nodes, and two authors are linked if they have collaborated to write a paper. Authors belonging to the same research area are more likely to collaborate, so the network has homophilic structure with research areas as communities. This gives a homogeneous network with 4,057 connected nodes and 3,528 links. Figure 4 (left column) shows a heavily right-skewed histogram and a roughly linear log empirical tail distribution plot, indicating that we should prefer Hom-RSC over Hom-SC for clustering this network.

However, 1466 of the 4057 authors have no edges in the author-author network and hence they have to be discarded, as disconnected nodes cannot be clustered either by Hom-SC or by Hom-RSC. We implemented the Hom-RSC algorithm from Section 4 on the remaining 2,591 nodes with $K = 4$ clusters. It turns out that 482 rows of the eigenvector matrix $\mathbf{X}$ are null rows which can not be normalized and hence cannot be clustered. After discarding them, we performed clustering on the remaining 2,109 author nodes. The algorithm misclustered 1,274 (60.41%) of these nodes. If we randomly assign the discarded nodes to the 4 clusters, the weighted average error rate for all 4,057 nodes is 67.41%. This number is the weighted average of clustering error (60.41%) and random assignment error (75%). We also implemented the Hom-SC algorithm — the error rate is 69.74% for the 2,591 connected authors and 71.64% for all authors. Hom-SC does not have a problem with clustering null rows in the eigenvector matrix. Thus, while Hom-RSC does perform better than Hom-SC, both homogeneous algorithms have accuracy comparable to random assignment to clusters.

7.2. Heterogenous author-paper-conference network

The DBLP system consists of *authors*, *papers*, and *conferences*. A heterogeneous network representation (APC network) of the DBLP system can thus be constructed with these three types of nodes and two types of links: author-paper (author writes paper) links and paper-conference (paper presented at conference) links. That authors are more likely to write papers in their research area, and papers are more likely to be presented at a conference belonging to the same research area, indicates homophilic community structure. This is a network with 18,405 nodes (4,057 authors, 14,328 papers, 20 conferences) and 33,973 links (19,645 author-paper links and 14,328 paper-conference links). All authors are now connected. The middle column of Figure 4 shows a heavily right-skewed histogram and a roughly linear log empirical tail distribution plot for author node degrees, indicating that we should prefer Het-RSC over Het-SC for clustering this network.

We implemented the Het-RSC algorithm on this network with $T = 3$ and $K = 4$. The error rate for authors was 7.30%. We also implemented Het-SC which gave an error rate of 23.10% for the authors. Thus, Het-RSC was quite accurate in identifying research area for authors from the heterogeneous network. Even Het-SC performed relatively well, although Het-RSC was more accurate than Het-SC as expected for a sparse network. In contrast, the homogeneous algorithms have accuracy similar to random allocation, which implies that the homogeneous co-authorship network is not very informative towards identification of authors' research area.

7.3. Heterogeneous author-conference network

The problem of interest in the four-area DBLP dataset is assigning authors to research communities, which is a homogeneous problem relating only to author nodes. However, the DBLP system itself is heterogeneous, and this heterogeneous information can be useful towards solving the homogeneous problem. In Section 7.2, we used data from the heterogeneous DBLP system to add two additional types of nodes (papers and conferences) to construct a heterogeneous network. This is the 'default' way to construct the heterogeneous network, using all the data at our disposal, and this approach gives us a much better solution to the problem than the homogeneous approach.

Suppose we instead consider a heterogeneous sub-system, and add only conference nodes, forming a smaller heterogeneous network with two types of nodes (authors and conferences) and only one type of link, author-conference (author presented at the conference). Authors from a research area are more likely to present at a conference related to the same area, indicating a homophilic community structure. This gives a network with 4,077 nodes (4,057 authors and 20 conferences) and 9,205 author-conference links. All authors are connected. The right column of Figure 4 shows a histogram that is right-skewed but not as heavily right-skewed as the two earlier networks. The log empirical tail distribution plot is also less linear than the other two networks. The node degrees vary between 1 and 14, which is a much tighter range than the degree range in the homogeneous author network or the heterogeneous APC network. Thus the network features do indicate sparsity, but less so than the two previous DBLP networks.

Implementing the Het-RSC algorithm on this bi-type network, we found an error rate of 7.44%, comparable to the error rate of Het-RSC in the APC network. The Het-SC algorithm gave an error rate of 8.85%, which is better than Het-SC in APC network. Both error rates are significant improvement over the homogeneous approach. Such improvement is achieved with only 20 additional nodes and therefore at a computational cost comparable to the homogeneous approach, while the APC network requires the addition of 14,348 nodes and therefore has greater computational cost.

The community detection problem of interest is often homogeneous, in the sense that it is defined with respect to only one type of agent, while the underlying system is heterogeneous. The user has the flexibility to choose from several heterogeneous sub-systems of the data to create a heterogeneous network. For example in the DBLP system, the user can choose the entire author-paper-conference system, or the author-conference subsystem, and so on. Consequently, the user might be interested in using an *optimal sub-system* that delivers the best community detection for the problem. One interesting avenue of future work is

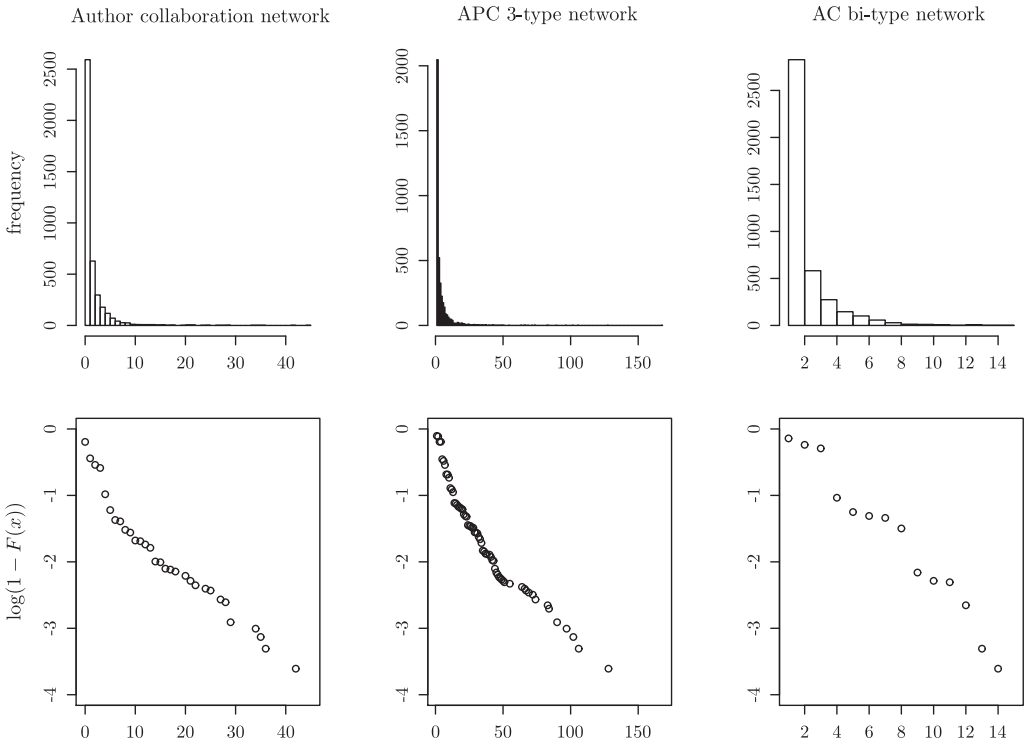


Figure 4. DBLP author degree distribution of homogeneous author collaboration network (left column), heterogeneous author-paper-conference network (middle column), and heterogeneous author-conference network (right column). Histograms (top row) of author node degrees have high frequency of low degrees, indicating that the author nodes are sparsely connected. The bottom row shows that the log empirical tail distributions $\log_{10}(1 - \hat{F}(x))$ are roughly linear, suggesting power-law behavior of author node degrees.

to lay down explicit criteria for selecting the optimal sub-network, akin to the analogous problem of variable selection in a machine learning framework.

8. Discussion

This paper introduces heterogeneous networks to the statistics literature, and extends the existing statistical framework of community detection in homogeneous networks to heterogeneous networks. We formulate heterogeneous versions of standard spectral clustering and regularized spectral clustering algorithms. The proposed algorithms have theoretical accuracy under heterogeneous versions of the SBM and the DCBM, respectively. Our simulations demonstrate that, even though homogeneous and heterogeneous methods have similar order of theoretical accuracy in large samples, the heterogeneous methods provide significantly better clustering results in finite-sample networks generated from several

interesting model settings. This comparative advantage seems to imply that the superiority of heterogeneous clustering over homogeneous clustering should be theoretically demonstrable, but we leave that to future work. The practical usefulness of the heterogeneous procedure is also demonstrated by the DBLP four-area dataset example, where the heterogeneous method delivers a far better clustering performance compared to the homogeneous method.

Acknowledgement

This work was supported in part by NSF grants DMS-11-06796 and DMS-14-06455. We thank the Editor, an associate editor, and two referees for their constructive suggestions and comments.

Appendix. Proof of Theorem 1

We prove the theorem for $T = 2$, bi-type heterogeneous networks. The proof easily generalizes to higher values of T. Consider a K-block, bi-type Het-DCBM with n_1 nodes of type 1 and n_2 nodes of type 2, and let $\tau \geq 0$ be the regularizer. Take $N = n_1 + n_2$, and let $\mathbf{X}_\tau, \mathcal{X}_\tau, \mathbf{X}_\tau^*$, and $\mathcal{X}_\tau^*$ be N-by-$2K$ matrices defined as per Sections 4 and 5.

Partition $\mathbf{X}_\tau^*$ as $\mathbf{X}_\tau^* = \begin{pmatrix} \mathbf{X}_\tau^{*(1)} \\ \mathbf{X}_\tau^{*(2)} \end{pmatrix}$, where $\mathbf{X}_\tau^{*(1)}$ is n_1-by-$2K$ and $\mathbf{X}_\tau^{*(2)}$ is n_2-by-$2K$. Then cluster centroids are

$$\mathbf{C}_t = \arg\min_{\mathbf{Y}_t \in \mathcal{Y}_t} ||\mathbf{X}_\tau^{*(t)} - \mathbf{Y}_t||_F^2 \quad \text{for} \quad t = 1, 2, \tag{A.1}$$

where $\mathcal{Y}_t = \{\mathbf{Y}_t \in \mathcal{R}^{n_t \times 2K} : \mathbf{Y}_t \text{ has } K \text{ unique rows}\}$, for $t = 1, 2$.

For the bi-type Het-DCBM, $\mathbf{M}$ has the form $\mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{22} \end{pmatrix}$, where $\mathbf{M}_{11}$ is n_1-by-K with exactly K distinct rows, and $\mathbf{M}_{22}$ is n_2-by-K with exactly K distinct rows. By Lemma 3.3 (2) of Qin and Rohe (2013), $\mathcal{X}_\tau^*$ can be expressed as $\mathcal{X}_\tau^* = \mathbf{MB}$ under the general DCBM, and hence also under the Het-DCBM, where $\mathbf{B}$ is a non-singular matrix of dimension $2K$-by-$2K$. Partition $\mathbf{B}$ into four K-by-K matrices as $\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}$. Then,

$$\mathcal{X}_\tau^* = \mathbf{MB} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{M}_{11}\mathbf{B}_{11} & \mathbf{M}_{11}\mathbf{B}_{12} \\ \mathbf{M}_{22}\mathbf{B}_{21} & \mathbf{M}_{22}\mathbf{B}_{22} \end{pmatrix} = \begin{pmatrix} \mathcal{X}_\tau^{*(1)} \\ \mathcal{X}_\tau^{*(2)} \end{pmatrix},$$

where $\mathcal{X}_\tau^{*(1)}$ is n_1-by-$2K$ and $\mathcal{X}_\tau^{*(2)}$ is n_2-by-$2K$. Since $\mathbf{M}_{11}$ and $\mathbf{M}_{22}$ have exactly K unique rows, and $\mathbf{B}$ is non-singular, $\mathcal{X}_\tau^{*(1)}$ and $\mathcal{X}_\tau^{*(2)}$ have K distinct rows: $\mathcal{X}_\tau^{*(1)} \in \mathcal{Y}_1$ and $\mathcal{X}_\tau^{*(2)} \in \mathcal{Y}_2$. Thus $\mathcal{X}_\tau^{*(t)}\mathbf{O} \in \mathcal{Y}_t$ for $t = 1, 2$, where $\mathbf{O}$ is an orthonormal rotation.

Without loss of generality we focus on $t = 1$. From the definition of $\mathbf{C}_1$ and the fact that $\mathcal{X}_\tau^{*(1)}\mathbf{O} \in \mathcal{Y}_1, ||\mathbf{X}_\tau^{*(1)} - \mathbf{C}_1||_F \leq ||\mathbf{X}_\tau^{*(1)} - \mathcal{X}_\tau^{*(1)}\mathbf{O}||_F$. So,

$$||\mathbf{C}_1 - \mathcal{X}_\tau^{*(1)}\mathbf{O}||_F \leq ||\mathbf{C}_1 - \mathbf{X}_\tau^{*(1)}||_F + ||\mathbf{X}_\tau^{*(1)} - \mathcal{X}_\tau^{*(1)}\mathbf{O}||_F \leq 2||\mathbf{X}_\tau^{*(1)} - \mathcal{X}_\tau^{*(1)}\mathbf{O}||_F.$$

Now

$$\mathcal{E}_1 = \{i \in \text{type 1} : \exists\ j \in \text{type 1}\ \&\ j \neq i\ s.t. \\ ||\mathbf{C}(i,\cdot) - \mathcal{X}_\tau^*(i,\cdot)\mathbf{O}||_2 > ||\mathbf{C}(i,\cdot) - \mathcal{X}_\tau^*(j,\cdot)\mathbf{O}||_2\}.$$

For two type 1 nodes $i \neq j$, either $M(i,\cdot) = M(j,\cdot)$ when they belong to the same block, or $M(i,\cdot)'M(j,\cdot) = 0$ when they belong to different blocks. Since $\mathcal{X}_\tau^* = \mathbf{MB}$, $\mathcal{X}_\tau^*$ is row-normalized, and $\mathbf{O}$ is orthonormal, so for two type 1 nodes $i \neq j$, either

$$\mathcal{X}_\tau^{*(1)}(i,\cdot)\mathbf{O} = \mathcal{X}_\tau^{*(1)}(j,\cdot)\mathbf{O} \Rightarrow ||\mathcal{X}_\tau^{*(1)}(i,\cdot)\mathbf{O} - \mathcal{X}_\tau^{*(1)}(j,\cdot)\mathbf{O}||_2 = 0$$

or

$$(\mathcal{X}_\tau^{*(1)}(i,\cdot)\mathbf{O})'(\mathcal{X}_\tau^{*(1)}(j,\cdot)\mathbf{O}) = 0 \Rightarrow ||\mathcal{X}_\tau^{*(1)}(i,\cdot)\mathbf{O} - \mathcal{X}_\tau^{*(1)}(j,\cdot)\mathbf{O}||_2 = \sqrt{2}.$$

This leads to the observation that

$$||\mathbf{C}(i,\cdot) - \mathcal{X}_\tau^{*(1)}(i,\cdot)\mathbf{O}||_2 < \frac{1}{\sqrt{2}} \Rightarrow \\ ||\mathbf{C}(i,\cdot) - \mathcal{X}_\tau^{*(1)}(i,\cdot)\mathbf{O}||_2 \leq ||\mathbf{C}(i,\cdot) - \mathcal{X}_\tau^{*(1)}(j,\cdot)\mathbf{O}||_2,\ \forall\ j \neq i.$$

which means $||\mathbf{C}(i,\cdot) - \mathcal{X}_\tau^{*(1)}(i,\cdot)\mathbf{O}||_2 < 1/\sqrt{2}$ is a sufficient condition for node i to be correctly clustered. Define $\mathcal{E}_1'$ to be the set of nodes that do not satisfy this sufficient condition, i.e.,

$$\mathcal{E}_1' = \Big\{i \in \text{type 1} : ||\mathbf{C}(i,\cdot) - \mathcal{X}_\tau^{*(1)}(i,\cdot)\mathbf{O}||_2 \geq \frac{1}{\sqrt{2}}\Big\}.$$

Then,

$$|\mathcal{E}_1| \leq |\mathcal{E}_1'| = \sum_{i \in \mathcal{E}_1'} 1 \leq 2 \sum_{i \in \mathcal{E}_1'} ||\mathbf{C}(i,\cdot) - \mathcal{X}_\tau^{*(1)}(i,\cdot)\mathbf{O}||_2^2 \leq 2||\mathbf{C}_1 - \mathcal{X}_\tau^{*(1)}\mathbf{O}||_F^2 \\ \leq 8||\mathbf{X}_\tau^{*(1)} - \mathcal{X}_\tau^{*(1)}\mathbf{O}||_F^2.$$

From (5.1), we have

$$||\mathbf{X}_\tau - \mathcal{X}_\tau\mathbf{O}||_F \leq c_0 \frac{1}{\lambda}\sqrt{\frac{K\log(4N/\epsilon)}{\delta + \tau}}$$

under (A1) and (A2). For any i,

$$||\mathbf{X}_\tau^*(i,) - \mathcal{X}_\tau^*(i,)\mathbf{O}||_2 \leq \frac{||\mathbf{X}_\tau(i,) - \mathcal{X}_\tau(i,)\mathbf{O}||_2}{\min\{||\mathbf{X}_\tau(i,)||_2, ||\mathcal{X}_\tau(i,)||_2\}}.$$

Therefore, from the definition of γ_1,

$$8||\mathbf{X}_\tau^{*(1)} - \mathcal{X}_\tau^{*(1)}\mathbf{O}||_F^2 \leq \frac{8||\mathbf{X}_\tau^{(1)} - \mathcal{X}_\tau^{(1)}\mathbf{O}||_F^2}{\gamma_1^2} \leq \frac{8||\mathbf{X}_\tau - \mathcal{X}_\tau\mathbf{O}||_F^2}{\gamma_1^2} \leq 8c_0^2 \frac{K\log(4N/\epsilon)}{\lambda^2\gamma_1^2(\delta+\tau)}.$$

This completes the proof for $T = 2$.

References

Amini, A. A., Chen, A., Bickel, P. J. and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41**, 2097-2122.

Chaudhuri, K., Chung, F. and Tsiatas, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. *J. Mach. Learn. Res.: Workshop and Conference Proceedings* **23**, 35.1-35.23.

Erdös, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen* **6**, 290-297.

Fienberg, S. E., Meyer, M. M. and Wasserman, S. S. (1985). Statistical analysis of multiple sociometric relations. *J. Amer. Statist. Assoc.* **80**, 51-67.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports* **486**, 75-174.

Gao, J., Liang, F., Fan, W., Sun, Y. and Han, J. (2009). Graph-based consensus maximization among multiple supervised and unsupervised models. *Adv. Neural Inform. Proc. Systems* **22**, 585-593.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M. and others (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147.

Girvan, M. and Newman, M. (2002). Community structure in social and biological networks. *Proc. Nat. Acad. Sci.* **99**, 7821-7826.

Goldenberg, A., Zheng, A. X., Fienberg, S. E. and Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning* **2**, 129-233.

Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983). Stochastic blockmodels: first steps. *Social Networks* **5**, 109-137.

Huberman, B. A. and Adamic, L. A. (1999). Internet: growth dynamics of the World-Wide Web. *Nature* **401**, 131-131.

Ji, M., Sun, Y., Danilevsky, M., Han, J. and Gao, J. (2010). Graph regularized transductive classification on heterogeneous information networks. *Machine Learning and Knowledge Discovery in Databases*, 570-586. Springer.

Jin, J. (2012). Fast network community detection by SCORE. arXiv:1211.5803.

Jonsson, P. F., Cavanna, T., Zicha, D. and Bates, P. A. (2006). Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics* **7**, 2.

Joseph, A. and Yu, B. (2013). Impact of regularization on spectral clustering. arXiv:1312.1733.

Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer.

Lorrain, F. and White, H. C. (1971). Structural equivalence of individuals in social networks. *J. Math. Soc.* **1**, 49-80.

Milgram, S. (1967). The small world problem. *Psychology Today*, **2**, 60-67.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press.

Qin, T. and Rohe, K. (2013). Regularized spectral clustering under the degreecorrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, 3120-3128.

Rohe, K., Chatterjee, S. and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39**, 1878-1915.

Sun, Y. and Han, J. (2012). *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan & Claypool Publishers.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**, 440-442.

Department of Statistics, University of Illinois at Urbana-Champaign, 725 South Wright Street, Champaign, IL 61820, U.S.A.

E-mail: ssengpt2@illinois.edu

Department of Statistics, University of Illinois at Urbana-Champaign, 725 South Wright Street, Champaign, IL 61820, U.S.A.

E-mail: yuguo@illinois.edu

(Received August 2013; accepted June 2014)