BIOMETRICS 64, 1–21 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

Logistic Regression Augmented Community Detection for Networks with Application in Identifying Autism-Related Gene Pathways

Yunpeng Zhao[1,*], Qing Pan[2,], and Chengan Du[1,***]**

[1]Department of Statistics, George Mason University, Fairfax Virginia 22030, U.S.A.

[2]Department of Statistics, George Washington University, Washington DC 20032, U.S.A.

**email:* yzhao15@gmu.edu

***email:* qpan@gwu.edu

****email:* cdu3@masonlive.gmu.edu

SUMMARY: When searching for gene pathways leading to specific disease outcomes, we propose to take advantage of additional information on gene characteristics to differentiate genes of interests from irrelevant background ones when connections involving both types of genes are observed and their relationships to the disease are unknown. Novel community detection methods are proposed that singles out irrelevant background genes with the help of auxiliary information through a logistic regression, and clusters relevant genes into cohesive groups using the adjacency matrix. Expectation-maximization algorithm is modified to maximize a joint pseudo-likelihood assuming latent indicators for relevance to the disease and latent group memberships as well as Poisson or multinomial distributed link numbers within and between groups. A robust version allowing arbitrary structures within the background is further derived. Asymptotic consistency of label assignments under the stochastic blockmodel is proven. Superior performance and robustness in finite samples are observed in simulation studies. The proposed robust method identifies previously missed gene sets underlying autism and related neurological diseases using diverse data sources including de novo mutations, gene expression and protein-protein interactions.

KEY WORDS: Autism spectral disease; Covariates augmented community detection; Expectation-maximization algorithm; Gene clustering; Pathway detection; Pseudo-likelihood.

1. Introduction

Community detection is a fundamental question in network analysis (Goldenberg et al., 2010; Newman, 2006; Fortunato, 2010). Traditional approaches consider the adjacency matrix, whose elements equal one or zero indicating whether there is a connection between two nodes, as the input. Then the nodes are partitioned into cohesive groups, that is, communities, with more links within and fewer links between the groups. Current community detection methods assume all nodes belong to certain communities of interests. However, this assumption is not always true in real applications. For example, when we are looking for pathways involving many genes that lead to certain disease, connections between candidate genes regardless their involvement in the disease process are collected. Furthermore, whether a gene is related to the disease is usually unknown. We propose to utilize information on the characteristics of the nodes/genes to differentiate between the nodes related to the outcome of interests and the unrelated ones. Novel two-stage models with one joint likelihood are proposed to incorporate the node-specific information which isolate relevant nodes from irrelevant ones and in return improve detection accuracy of communities related to a specific outcome.

Our study is motivated by the problem to discover gene pathways leading to complex diseases in genomic studies. Multiple sources of data, e.g. highly correlated gene expression levels and experimentally verified protein-protein interactions, provide useful information on connections between genes. However, not all genes are related to the disease under study. In fact, most genes are “household” genes that function to maintain the normal metabolic processes within healthy human bodies. Mixing genes and pathways for normal life processes with those leading to the target disease in community detection models will introduce noise as well as impurity to disease-generating pathways which are the true interests of clinicians and biologists. De novo mutations refer to gene mutations that occur for the first time in a family compared to mutations inherited from parents. We believe that discrepancy in the numbers of de novo mutations on the same gene in cases and controls would help differentiate genes related to the disease from those unrelated to the disease, which we call the “background”. The proposed novel method is feasible because the three kinds of data, gene expression, protein-protein interaction and number of de novo mutations, can be downloaded from different online data consortiums and combined using unique gene names. In summary, our method targets at gene groups with the following characteristics – 1) cases have a higher frequency of de novo mutations than controls, 2) concurrent expression patterns within the same group, and 3) dense protein-protein interactions within the same module and sparse interactions between different groups.

The stochastic blockmodel is the most used statistical tool for modeling and detecting communities (Holland et al., 1983; Snijders and Nowicki, 1997; Nowicki and Snijders, 2001). We generalize the blockmodel by modeling the relationship between the unobserved indicator whether a gene is related to the target disease or not and gene-specific covariates in the first stage, then cluster disease-related genes into closely connected pathways in the second stage. Because both indicators for disease relevance in the first stage and community labels in the second stage are latent variables, the expectation-maximization algorithm is employed. However, this approach is intractable due to the numerous possible label assignments in the E-step. Amini et al. (2013) proposed a fast pseudo-likelihood algorithm for fitting blockmodels and we adapt this algorithm in Section 3 to the joint pseudo-likelihoods incorporating both the logistic regression and the block models. The pseudo-likelihood may also be optimized by other alternative approaches such as the EMM algorithm by Gormley and Murphy (2008).

Another distinct feature of the proposed method is the extension to the robust community detection allowing heterogeneous linkage probabilities in the background, which relaxes the assumption of homogeneous linkage probability within each group in the stochastic blockmodel. For instance, the background can be a mixture of multiple strongly or weakly connected groups. These groups all belong to the background because they are not related to the target disease, but their structure is not necessarily homogeneous. In Section 4, we further develop the model in section 3 to allow for arbitrary structures within the background. Interestingly, when the linkage probabilities within the background are unspecified, the pseudo-likelihood algorithm can be easily adapted to leave the likelihood of the links in the background out while the classical likelihood approach cannot.

Recently there have been works on community detection which utilize covariates information. These papers are seeking how to use the additional covariates information to improve the accuracy of community detection. This task is sometimes achieved by combining a similarity or kernel matrix defined based on covariates with the adjacency matrix (Binkiewicz et al., 2014; Zhang et al., 2015; Yan and Sarkar, 2016; Xu et al., 2012). On the other side, likelihoods of linkage probabilities incorporating auxiliary nodal information have been proposed by Tallberg (2004), Yang et al. (2013), Newman and Clauset (2016), Handcock et al. (2007), Krivitsky et al. (2009) and Gormley and Murphy (2010). However, none of these works follow the same framework as our method. In short, the sole reason of using auxiliary information on nodal characteristics in our method is to distinguish the disease related nodes from unrelated ones, then we carry out community detection within the disease-related nodes. On the contrary, auxiliary information in the literature is usually used to facilitate partition of all nodes into communities. For example, Tallberg (2004) used covariates to predict the probabilities into each homogeneous community in a Bayesian framework, while we use covariates to predict the probability into the heterogeneous background in a pseudo-likelihood framework.

2. Methods

We begin by introducing the data structure and notation. A network with n nodes can be represented by an $n \times n$ adjacency matrix $A = [A_{ij}]$, where

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between } i \text{ and } j, \\ 0 & \text{otherwise} \end{cases}$$

In addition to the adjacency matrix A, some covariate information on nodes is also available. These covariates are represented by an $n \times P$ matrix $X = [x_{ip}]$, where x_{ip} denotes the value of the pth covariate on node i.

We model networks with a particular community structure where the network is composed of multiple cohesive communities, together with some *background* nodes. Unlike the usual definition of background set which is diffuse within itself or weakly connected to other parts of the network (Zhao et al., 2011), we assume that the probability of a node belonging to the background set depends on its covariates. Suppose there are K communities besides the background set. Let $\boldsymbol{c} = (c_1, c_2, ..., c_n)$ denote the community that each of the n nodes/genes belongs to, thus $c_i = k$ if nodes i belongs to community k, for $k \in \{1, 2, ..., K\}$, and $c_i = K + 1$ if node i is a background gene. Moreover, let $\boldsymbol{y} = [y_i]$ be a vector indicating whether the node belongs to one of the K communities or the background, i.e. $y_i = 1$ if $c_i \leqslant K$, $y_i = 0$ otherwise.

The network is generated in three steps.

STEP 1: The random variable y_i is independent for $i = 1, \cdots, n$ and follows a logistic regression

$$\text{pr}(y_i = 1 \mid X) = \frac{e^{\boldsymbol{x}_i \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i \boldsymbol{\beta}}},$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_P)^T$ is the coefficients vector, and $\boldsymbol{x}_i$ is the ith row of X. Here the logistic model has an intercept, that is, X contains $(1, 1, ..., 1)^T$ as its first column.

STEP 2: The probability that a node with $y_i = 1$ belongs each of the K communities is given by the independent multinomial distribution with parameter $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$,

$$\text{pr}(c_i = k \mid y_i = 1) = \pi_k, \quad (i = 1, ..., n; k = 1, ..., K).$$

In addition, $c_i = K + 1$ if $y_i = 0$.

STEP 3: Conditional on the labels, A_{ij} for $i < j$ are independent Bernoulli variables with

$$\mathrm{pr}(A_{ij} = 1 \mid \boldsymbol{c}) = P_{c_i c_j},$$

where P is a $(K + 1) \times (K + 1)$ symmetric matrix.

The total number of genes in the kth community is $n_k = \sum_{i=1}^n 1(c_i = k)$ and the number of links between the kth and lth commuity is given by $O_{kl} = \sum_{1 \leqslant i,j \leqslant n} A_{ij} 1(c_i = k, c_j = l)$, where $1(\cdot)$ is the indicator function. Moreover, let $n_{kl} = n_k n_l$ if $k \neq l$, and $n_{kk} = n_k(n_k - 1)$. Then the joint log-likelihood of $\boldsymbol{c}$ and A is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\pi}, P; \boldsymbol{c}, A) = &\sum_{i=1}^n \{y_i \boldsymbol{x}_i \boldsymbol{\beta} - \log(1 + e^{\boldsymbol{x}_i \boldsymbol{\beta}})\} + \sum_{k=1}^K n_k \log \pi_k \\ &+ \frac{1}{2} \sum_{1 \leqslant k,l \leqslant K+1} \{O_{kl} \log P_{kl} + (n_{kl} - O_{kl}) \log(1 - P_{kl})\}.\end{aligned}$$

3. Estimating Procedures

The community labels $\boldsymbol{c}$ are unobserved in a community detection problem. Furthermore, the E-step of such algorithm requires evaluating all the possible label assignments, which makes the algorithm intractable (Amini et al., 2013; Zhao et al., 2012). We adopt the idea of pseudo-likelihood in Amini et al. (2013) which partitions each row of A into blocks and assumes the independence among rows.

We briefly review some notation used in Amini et al. (2013). The vector $\boldsymbol{e} = (e_1, ..., e_n)$ denotes an initial blocking vector, where $e_i \in \{1, ..., K + 1\}$. And b_{ik} denotes the number of edges associated with node i in the kth block, that is, $b_{ik} = \sum_{j=1}^n A_{ij} 1(e_i = k)$ $(i = 1, .., n; j = 1, ..., K + 1)$. Let $B = [b_{ik}]_{1 \leqslant i \leqslant n, 1 \leqslant k \leqslant K+1}$ and $\Lambda = [\lambda_{lk}]_{1 \leqslant l,k \leqslant K+1}$, where λ_{lk} is the expected total number of edges in the k-th block for a node i in community l, i.e., $c_i = l$. When n is large, b_{ik} can be approximated by a Poisson distribution given c_i, and the dependence of B between different rows is weak. Assuming b_{ik} are independence for $i = 1, \cdots, n$ and $k = 1, \cdots, K + 1$ and using the

Poisson approximation, the log-pseudolikelihood of $\boldsymbol{c}$ and B (up to a constant) is

$$\sum_{i=1}^{n}\{y_i\boldsymbol{x}_i\boldsymbol{\beta} - \log(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}})\} + \sum_{k=1}^{K} n_k \log\pi_k + \sum_{i=1}^{n}\sum_{l=1}^{K+1} 1(c_i = l)\left(-\mu_l + \sum_{k=1}^{K+1} b_{ik}\log\lambda_{lk}\right),$$

where $\mu_l = \sum_k \lambda_{lk}\ (l = 1, ..., K+1)$. And the log of marginal distribution of B (up to a constant) is

$$\begin{aligned}\mathcal{L}_{\text{Poisson}}(\boldsymbol{\beta}, \boldsymbol{\pi}, \Lambda; B) = &\sum_{i=1}^{n}\log\left\{\sum_{l=1}^{K}\frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}}\pi_l e^{-\mu_l}\left(\prod_{k=1}^{K+1}\lambda_{lk}^{b_{ik}}\right)\right.\\ &\left. + \frac{1}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}}e^{-\mu_{K+1}}\left(\prod_{k=1}^{K+1}\lambda_{K+1,k}^{b_{ik}}\right)\right\}. \qquad (1)\end{aligned}$$

Given initial labels $\boldsymbol{e}$, equation (1) can be maximized by a standard expectation-maximization algorithm. The details of the E-step and M-step are given in Algorithm 1.

Algorithm 1: (The expectation-maximization algorithm under Poisson distribution)

- E-step: Let $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\pi}}$ and $\hat{\Lambda}$ be the estimates at the current iteration, and $\hat{\mu}_l = \sum_k \hat{\lambda}_{lk}\ (l = 1, ..., K+1)$. The posterior probability of label assignment is

$$\begin{aligned} z_{il} &= \text{pr}(c_i = l \mid B)\\ &= \frac{\frac{e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}{1+e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}\hat{\pi}_l e^{-\hat{\mu}_l}\left(\prod_{k=1}^{K+1}\hat{\lambda}_{lk}^{b_{ik}}\right)}{\sum_{l=1}^{K}\frac{e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}{1+e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}\hat{\pi}_l e^{-\hat{\mu}_l}\left(\prod_{k=1}^{K+1}\hat{\lambda}_{lk}^{b_{ik}}\right) + \frac{1}{1+e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}e^{-\hat{\mu}_{K+1}}\left(\prod_{k=1}^{K+1}\hat{\lambda}_{K+1,k}^{b_{ik}}\right)}\\ &\qquad (i = 1, ..., n; l = 1, ..., K),\end{aligned}$$

$$\begin{aligned} z_{i,K+1} &= \text{pr}(c_i = K+1 \mid B)\\ &= \frac{\frac{1}{1+e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}e^{-\hat{\mu}_{K+1}}\left(\prod_{k=1}^{K+1}\hat{\lambda}_{K+1,k}^{b_{ik}}\right)}{\sum_{l=1}^{K}\frac{e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}{1+e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}\hat{\pi}_l e^{-\hat{\mu}_l}\left(\prod_{k=1}^{K+1}\hat{\lambda}_{lk}^{b_{ik}}\right) + \frac{1}{1+e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}e^{-\hat{\mu}_{K+1}}\left(\prod_{k=1}^{K+1}\hat{\lambda}_{K+1,k}^{b_{ik}}\right)}\\ &\qquad (i = 1, ..., n).\end{aligned}$$

- M-step: Given $z_{il}\ (i = 1, ...n; l = 1, ..., K+1)$, $\hat{\pi}$ and $\hat{\Lambda}$ can be updated by closed form formulae,

$$\begin{aligned}\hat{\pi}_l &= \frac{\sum_i z_{il}}{\sum_i\sum_{l=1}^{K} z_{il}} \quad (l = 1, ..., K),\\ \hat{\lambda}_{lk} &= \frac{\sum_i z_{il}b_{ik}}{\sum_i z_{il}} \quad (l = 1, ..., K+1; k = 1, ..., K+1).\end{aligned}$$

$\hat{\boldsymbol{\beta}}$ can be updated by logistic regression,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{i=1}^{n} \left\{ \left(\sum_{l=1}^{K} z_{il} \right) \boldsymbol{x}_i \boldsymbol{\beta} - \log(1 + e^{\boldsymbol{x}_i \boldsymbol{\beta}}) \right\}.$$

Note $\sum_{l=1}^{K} z_{il}$ is the sum of the estimated conditional probabilities of gene i belonging to one of the K communities.

Once the expectation-maximization algorithm converges, we can update the labels $\boldsymbol{e}$ by $e_i = \operatorname{argmax}_{1 \leqslant l \leqslant K+1} z_{il}$. We repeat this procedure several times until $\boldsymbol{e}$ becomes stable.

Amini et al. (2013) introduced a pseudo-likelihood conditional on the node degrees. We generalize this conditional pseudo-likelihood to our scenario. Denote the node degree by $d_i = \sum_k b_{ik}$ $(i = 1, ..., n)$. Then $(b_{i1}, ..., b_{i,K+1})$ follows multinomial distribution conditional on label $\boldsymbol{c}$ and d_i. The multinomial log pseudo-likelihood (up to a constant) is

$$\begin{aligned} \mathcal{L}_{\text{Multinomial}}(\boldsymbol{\beta}, \boldsymbol{\pi}, \Theta; B) = & \sum_{i=1}^{n} \log \left\{ \sum_{l=1}^{K} \frac{e^{\boldsymbol{x}_i \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i \boldsymbol{\beta}}} \pi_l \left(\prod_{k=1}^{K+1} \theta_{lk}^{b_{ik}} \right) \right. \\ & \left. + \frac{1}{1 + e^{\boldsymbol{x}_i \boldsymbol{\beta}}} \left(\prod_{k=1}^{K+1} \theta_{K+1,k}^{b_{ik}} \right) \right\}, \end{aligned} \tag{2}$$

where $\Theta = [\theta_{lk}]$ $(l = 1, ..., K+1; k = 1, ..., K+1)$ is the parameter in the multimomial distribution satisfying $\sum_{k=1}^{K+1} \theta_{lk} = 1 (l = 1, ..., K+1)$.

The algorithm is similar to that for the Poisson pseudo-likelihood. For completeness, we give the details of the expectation-maximization algorithm under the multinomial distribution in Algorithm 2.

Algorithm 2: (The expectation-maximization algorithm under multinomial distribution)

- E-step: Based on current estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\pi}}$ and $\hat{\Theta}$, the posterior probability of label assignment is

$$z_{il} = \frac{\frac{e^{\boldsymbol{x}_i \hat{\boldsymbol{\beta}}}}{1+e^{\boldsymbol{x}_i \hat{\boldsymbol{\beta}}}} \hat{\pi}_l \left(\prod_{k=1}^{K+1} \hat{\theta}_{lk}^{b_{ik}} \right)}{\sum_{l=1}^{K} \frac{e^{\boldsymbol{x}_i \hat{\boldsymbol{\beta}}}}{1+e^{\boldsymbol{x}_i \hat{\boldsymbol{\beta}}}} \hat{\pi}_l \left(\prod_{k=1}^{K+1} \hat{\theta}_{lk}^{b_{ik}} \right) + \frac{1}{1+e^{\boldsymbol{x}_i \hat{\boldsymbol{\beta}}}} \left(\prod_{k=1}^{K+1} \hat{\theta}_{K+1,k}^{b_{ik}} \right)} \quad (i = 1, ..., n; l = 1, ..., K),$$

$$z_{i,K+1} = \frac{\frac{1}{1+e^{\boldsymbol{x}_i \hat{\boldsymbol{\beta}}}} \left(\prod_{k=1}^{K+1} \hat{\theta}_{K+1,k}^{b_{ik}} \right)}{\sum_{l=1}^{K} \frac{e^{\boldsymbol{x}_i \hat{\boldsymbol{\beta}}}}{1+e^{\boldsymbol{x}_i \hat{\boldsymbol{\beta}}}} \hat{\pi}_l \left(\prod_{k=1}^{K+1} \hat{\theta}_{lk}^{b_{ik}} \right) + \frac{1}{1+e^{\boldsymbol{x}_i \hat{\boldsymbol{\beta}}}} \left(\prod_{k=1}^{K+1} \hat{\theta}_{K+1,k}^{b_{ik}} \right)} \quad (i = 1, ..., n).$$

- M-step: Given z_{il} $(i = 1, ...n; l = 1, ..., K + 1)$, $\hat{\pi}$, $\hat{\Theta}$ and $\hat{\boldsymbol{\beta}}$ can be updated by

$$\hat{\pi}_l = \frac{\sum_i z_{il}}{\sum_i \sum_{l=1}^{K} z_{il}} \quad (l = 1, ..., K),$$
$$\hat{\theta}_{lk} = \frac{\sum_i z_{il} b_{ik}}{\sum_i z_{il} d_i} \quad (l = 1, ..., K+1; k = 1, ..., K + 1),$$
$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{i=1}^{n} \left\{ \left(\sum_{l=1}^{K} z_{il} \right) \boldsymbol{x}_i \boldsymbol{\beta} - \log(1 + e^{\boldsymbol{x}_i \boldsymbol{\beta}}) \right\}.$$

4. Robust Community Detection

So far we assume that all the disease-related communities and the background satisfy the stochastic blockmodel assumption. In this section, we propose a new pseudo-likelihood method that allows for arbitrary structure in the background, for example, a mixture of tightly and weakly connected groups, or nodes with high degree variations. In other words, we still assume that the disease-related communities follow the stochastic blockmodel assumption, but make no assumption on structure within the background. As in Section 2, a network with the robust background is generated by three steps. The first two steps remain unchanged and the last step has been modified as follows.

STEP 3*: Conditional on the labels, when $k \leqslant K$ or $l \leqslant K$, A_{ij} for $i < j$ are independent Bernoulli variables with

$$\text{pr}(A_{ij} = 1 \mid c_i = k, c_j = l) = P_{kl}.$$

The link probabilities within the background set, i.e., when $k = K + 1$ and $l = K + 1$, are not specified.

It would not be helpful to consider the likelihood function contributed by the links within the background because part of the link probabilities are unspecified. By contrast, the pseudo-likelihood method introduced in Section 3 can be extended to this new scenario and provides interesting insights. Recall the setup in Section 3. Let $\boldsymbol{e} = (e_1, ..., e_n)$ be an initial blocking vector. And b_{ik} denotes the number of edges associated with node i in the kth block $(i = 1, .., n; j =$

$1, ..., K+1)$. Imagining that $\boldsymbol{e}$ is a reasonable initial vector, b_{ik} can be approximated by a mixture of Poisson distributions as before when $k = 1, ..., K$. But when $k = K+1$, the distribution of b_{ik} is unknown since the link probabilities within the background are unspecified. By excluding this part of unreliable information, we propose the following pseudo-likelihood for robust community detection,

$$\mathcal{L}_{\text{Robust}}(\boldsymbol{\beta}, \boldsymbol{\pi}, \Lambda; B, \boldsymbol{c}) = \sum_{i=1}^{n} \{y_i \boldsymbol{x}_i \boldsymbol{\beta} - \log(1 + e^{\boldsymbol{x}_i \boldsymbol{\beta}})\} + \sum_{k=1}^{K} n_k \log \pi_k$$
$$+ \sum_{i=1}^{n} \sum_{l=1}^{K+1} 1(c_i = l) \left(-\mu_l + \sum_{k=1}^{K} b_{ik} \log \lambda_{lk} \right), \qquad (3)$$

where $\mu_l = \sum_{k=1}^{K} \lambda_{lk}$ $(k = 1, .., K)$.

Notice equation (3) is indeed a valid likelihood function for a fixed $\boldsymbol{e}$ because the blocking vector $\boldsymbol{e}$ and the community labeling vector $\boldsymbol{c}$ are different. The blocking vector $\boldsymbol{e}$ partitions the columns of A into $K+1$ blocks and b_{ik} is the kth block sum for row i. Likelihood (3) does not include $B_{\cdot, K+1}$ - the last column of B since the Poisson approxmiation may not be valid. But this does not affect the range of c_i, which is still $\{1, ..., K+1\}$. Community detection based on (3) can be viewed as a classic clustering problem on B. We need to assign a label from 1 to $K+1$ to each row data point, i.e., each $B_{i\cdot}$, which contains $K+1$ features. But we only use the first K features since the last one is not reliable. The algorithm is therefore similar to Algorithm 1. For completeness, we give the details.

Algorithm 3: (The expectation-maximization algorithm for robust community detection)

- E-step: Let $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}$ and $\hat{\Lambda}$ be the estimates at the current iteration, and $\hat{\mu}_l = \sum_{k=1}^{K} \hat{\lambda}_{lk}$ $(l =$

$1, ..., K+1$). The posterior probability of label assignment is

$$\begin{aligned} z_{il} &= \mathrm{pr}(c_i = l \mid B) \\ &= \frac{\frac{e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}{1+e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}\hat{\pi}_l e^{-\hat{\mu}_l}\left(\prod_{k=1}^K \hat{\lambda}_{lk}^{b_{ik}}\right)}{\sum_{l=1}^K \frac{e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}{1+e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}\hat{\pi}_l e^{-\hat{\mu}_l}\left(\prod_{k=1}^K \hat{\lambda}_{lk}^{b_{ik}}\right) + \frac{1}{1+e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}} e^{-\hat{\mu}_{K+1}}\left(\prod_{k=1}^K \hat{\lambda}_{K+1,k}^{b_{ik}}\right)} \\ &\qquad\qquad\qquad\qquad\qquad\qquad (i = 1, ..., n; l = 1, ..., K), \end{aligned}$$

$$\begin{aligned} z_{i,K+1} &= \mathrm{pr}(c_i = K+1 \mid B) \\ &= \frac{\frac{1}{1+e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}} e^{-\hat{\mu}_{K+1}}\left(\prod_{k=1}^K \hat{\lambda}_{K+1,k}^{b_{ik}}\right)}{\sum_{l=1}^K \frac{e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}{1+e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}}\hat{\pi}_l e^{-\hat{\mu}_l}\left(\prod_{k=1}^K \hat{\lambda}_{lk}^{b_{ik}}\right) + \frac{1}{1+e^{\boldsymbol{x}_i\hat{\boldsymbol{\beta}}}} e^{-\hat{\mu}_{K+1}}\left(\prod_{k=1}^K \hat{\lambda}_{K+1,k}^{b_{ik}}\right)} \\ &\qquad\qquad\qquad\qquad\qquad\qquad (i = 1, ..., n). \end{aligned}$$

- M-step: Given z_{il} $(i = 1, ...n; l = 1, ..., K+1)$, $\hat{\pi}$, $\hat{\Lambda}$ and $\hat{\boldsymbol{\beta}}$ can be updated by,

$$\begin{aligned} \hat{\pi}_l &= \frac{\sum_i z_{il}}{\sum_i \sum_{l=1}^K z_{il}} \quad (l = 1, ..., K), \\ \hat{\lambda}_{lk} &= \frac{\sum_i z_{il} b_{ik}}{\sum_i z_{il}} \quad (l = 1, ..., K+1; k = 1, ..., K), \\ \hat{\boldsymbol{\beta}} &= \operatorname*{argmax}_{\boldsymbol{\beta}} \sum_{i=1}^n \left\{ \left(\sum_{l=1}^K z_{il} \right) \boldsymbol{x}_i\boldsymbol{\beta} - \log(1 + e^{\boldsymbol{x}_i\boldsymbol{\beta}}) \right\}. \end{aligned}$$

As before, once the expectation-maximization algorithm converges, $\boldsymbol{e}$ is updated by $e_i = \mathrm{argmax}_{1\leqslant l \leqslant K+1} z_{il}$. We repeat this procedure until $\boldsymbol{e}$ becomes stable.

We do not consider robust community detection using multinomial approximation since the condition $\sum_{k=1}^{K+1} \theta_{lk} = 1(l = 1, ..., K+1)$ becomes invalid if the last column is removed.

5. Asymptotic Properties

In this section we study the consistency under stochastic blockmodels. Equation (2) has slightly simpler form and theoretical derivations than (1). The theoretical analysis in this section will focus on the multinomial pseudo-likelihood.

We begin with the setup, which closely follow those in Amini et al. (2013). The case of one community with the background is taken as an example. The true community labels $\boldsymbol{c}$ are the parameters of interests, where $\pi_k = 1/n \sum_i 1(c_i = k)$ $(k = 1, 2)$. We focus on the case of directed blockmodel. A coupling technique can be used to extend the result to the undirected case analogous to that in Amini et al. (2013). Consider the edge matrix

$$P = \frac{1}{n}\begin{pmatrix} a_1 & b \\ b & a_2 \end{pmatrix} = \frac{b}{n}\begin{pmatrix} \rho_1 & 1 \\ 1 & \rho_2 \end{pmatrix},$$

where $\rho_k = a_k/b$. Here ρ_1 and ρ_2 remain constant, while b can scale with n. The directed blockmodel assumes that all the entries in the adjacency matrix are independent Bernoulli variables without forcing P to be symmetric, that is, $A_{ij} \sim$ Bernoulli$(P_{c_i c_j})$ $(i = 1, ..., n; j = 1, ..., n)$. For simplicity, a univariate covariate $\boldsymbol{x}$ taking values in $(1/n, 2/n, ..., 1)$ is assumed.

We illustrate the consistency of one-step expectation-maximization of the multinomial pseudo-likelihood. Starting from some initial labels $\boldsymbol{e}$ and initial estimates $\hat{b}$, $\hat{\rho}_1, \hat{\rho}_2$ of the parameters b, ρ_1 and ρ_2, the initial estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained from the logistic regression, that is,

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0,\beta_1}{\operatorname{argmax}} \sum_{i=1}^{n} \left\{ y_i(\beta_0 + x_i\beta_1) - \log(1 + e^{\beta_0 + x_i\beta_1}) \right\}.$$

Define

$$\hat{\pi}_{i1} = \frac{e^{\hat{\beta}_0 + x_i\hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + x_i\hat{\beta}_1}} \quad (i = 1, ..., n),$$
$$\hat{\pi}_{i2} = \frac{1}{1 + e^{\hat{\beta}_0 + x_i\hat{\beta}_1}} \quad (i = 1, ..., n).$$

Let

$$\hat{P} = \frac{\hat{b}}{n}\begin{pmatrix} \hat{\rho}_1 & 1 \\ 1 & \hat{\rho}_2 \end{pmatrix},$$

and R be the 2 by 2 matrix with entries $\{R_{ka}\}$ given by $R_{ka} = (1/n)\sum_{i=1}^{n} 1(e_i = k, c_i = a)$. The initial estimates $\hat{\Theta}$ is obtained by row normalization of $\hat{\Lambda} = [nR\hat{P}]^T$, that is,

$$\hat{\Theta} = \begin{pmatrix} \frac{\hat{\lambda}_{11}}{\hat{\lambda}_{11}+\hat{\lambda}_{12}} & \frac{\hat{\lambda}_{12}}{\hat{\lambda}_{11}+\hat{\lambda}_{12}} \\ \frac{\hat{\lambda}_{21}}{\hat{\lambda}_{21}+\hat{\lambda}_{22}} & \frac{\hat{\lambda}_{22}}{\hat{\lambda}_{21}+\hat{\lambda}_{22}} \end{pmatrix}.$$

With the notation defined above, the output of one-step expectation-maximization is

$$\hat{c}_i(\boldsymbol{e}) = \underset{k\in\{1,2\}}{\operatorname{argmax}} \left(\log\hat{\pi}_{ik} + \sum_{l=1}^{2} b_{il}\log\hat{\theta}_{kl} \right) \quad (i = 1, ..., n).$$

We use the mis-classification error rate (Choi et al., 2012; Zhao et al., 2012; Amini et al., 2013) to measure the performance of $\hat{c}_i$. That is, define

$$M_n(\boldsymbol{e}) = \min_{\phi \in \{(12),(21)\}} \frac{1}{n} \sum_{i=1}^{n} 1\{\hat{c}_i(\boldsymbol{e}) \neq \phi(c_i)\},$$

where $\{(12),(21)\}$ is the set of permutations of $\{1,2\}$. In this definition we consider all ϕ values that are permutations of each other because they result in the same community structure.

Consider the class of initial labels that correctly classify the node i as a member of community k. The fraction of such nodes among all nodes belonging to community k, γ_k, is formally given by

$$\mathcal{E} = \{\boldsymbol{e} : \sum_i 1(e_i = k, c_i = k) = \gamma_k n_k, k = 1, 2\},$$

where $n_k = \sum\limits_i 1(c_i = k)$ is the size of community k.

An extra condition is introduced to avoid perfect separation of $\boldsymbol{e}$ in the logistic fit. We define the following class

$$\mathcal{F} = \{\boldsymbol{e} : \sum_{i=\hat{n}_2+1}^{n} 1(e_i = 1) \leqslant \hat{n}_1 \tilde{\gamma}_1, \sum_{i=1}^{\hat{n}_1} 1(e_i = 1) \leqslant \hat{n}_1 \tilde{\gamma}_2\},$$

where $\hat{n}_k = \sum\limits_i 1(e_i = k)$ is the size of initial estimate of community k.

The uniform consistency of $\hat{c}_i$ within the class $\mathcal{E} \cap \mathcal{F}$ is established by the following theorem.

THEOREM 1 (Main result): *Assume* $\gamma_1, \gamma_2 \neq 1/2$ *and* $0 < \tilde{\gamma}_1, \tilde{\gamma}_2 < 1$. *Then under some regularity condition, with sufficiently large* $\hat{\rho}_1$, $\hat{\rho}_2$ *and* $b \to \infty$, *for any* ϵ,

$$\text{pr}\left[\sup_{\boldsymbol{e} \in \mathcal{E} \cap \mathcal{F}} M_n(\boldsymbol{e}) > \epsilon\right] \to 0, \quad \textit{as } n \to \infty.$$

The details of the regularity condition and the proof is given in the supplementary material.

The proof of the main theorem depends on a key fact that the log ratio of the estimated probabilities $\hat{\pi}_{i1}$ and $\hat{\pi}_{i2}$ has a uniform bound independent with n, for $\boldsymbol{e} \in \mathcal{F} \cap \mathcal{E}$. This is summarized in the following lemma.

LEMMA 1: *Assume* $0 < \tilde{\gamma}_1, \tilde{\gamma}_2 < 1$. *Then if* $\mathbf{e} \in \mathcal{F} \cap \mathcal{E}$, *there exist* M *such that for sufficiently* *large n,*

$$\left|\log \frac{\hat{\pi}_{i1}}{\hat{\pi}_{i2}}\right| < M,$$

where M is independent with n.

The proof is given in the supplementary material.

6. Simulations

We first examine the performance of the proposed methods under standard stochastic blockmodel. Each network contains $n = 500$ nodes and each setup is repeated 500 times. There are three groups including two disease-related communities and one disease-irrelevant background set. The probability a gene is related to the disease follows a logistic regression with logit pr$(y_i = 1 \mid x_i) = 4x_i + \beta_0$. Here y_i is the indicator for the ith node belonging to a disease-related community and covariate $x_i \sim U(-1, 1)$. And $\beta_0 = -1, 0, 1$ correspond to the percentages background 62%, 50% and 38%, respectively. Nodes with $y_i = 1$ are assigned to two non-overlapping communities with equal probabilities $\pi_1 = \pi_2 = 1/2$. Pairs within the background, as well as pairs composed of one node in the background and the other node in a disease-related community are linked with probability $0{\cdot}1$. The linkage probability between the two non-background communities is 0·05, while the linkage probability for pairs within the same community ranges from 0·15 to 0·25.

[Table 1 about here.]

Table 1 compares the performance of three models - the pseudo-likelihood methods with Poisson and multinomial approximation introduced in Section 3 as well as the robust community detection method introduced in Section 4. For each model, we further compare the two versions where auxiliary nodal information, i.e, logistic regression is either used or unused. The community detection accuracy is measured by the adjusted rand index (ARI) (Vinh et al., 2010). The performance of all methods improves as the linkage probability within disease-related community increases, or as the percentage of background nodes decreases. More importantly, the proposed method incorporating auxiliary information through logistic regression always outperforms the corresponding method without logistic regression. Moreover, the robust method gives the same performance as the Poisson pseudo-likelihood which suggests the robust method does not lose discriminatory accuracy when data follow standard stochastic block models. On the other hand, the algorithm fitting multinomial distributions performs slightly worse than the other two methods. Rigorously speaking, the multinomial pseudo-likelihood is an approximation to the degree-corrected blockmodel, which is a generalization of standard blockmodel by allowing more variation on degrees (Zhao et al., 2012; Karrer and Newman, 2011; Amini et al., 2013). Therefore, the finite sample performance of multinomial pseudo-likelihood has slightly lower ARI on average since it fits a more complicated model.

Next we consider the setup with heterogeneous background nodes. For any node i in background, we generate u_i from $U(0, 0{\cdot}2)$. The linkage probability between a background node i and a disease-related node is u_i. For two background nodes i and j, the linkage probability is $\sqrt{u_i u_j}$. The rest of the model setups such as the generation mechanism of communities labels, the linkage probabilities within/between communities and linkage probabilities between a community and the background remain the same.

[Table 2 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

The ARI of the six methods are shown in Table 2 and Figures 1 - 3. Similar to what we observed in Table 1, the average ARIs of all methods increases as the linkage probability within community increases, or as the percentage of background nodes decreases. And the method with logistic regression outperforms the corresponding method without logistic regression. The robust method with logistic regression gives the best performance in most scenarios. The Poisson pseudo-likelihood has the worst performance when the stochastic blockmodel assumption is violated in the heterogeneous background. Especially, under the case of high percentage of background nodes, the Poisson pseudo-likelihood performs poorly even when the linkage probability within community is high. The multinomial pseudo-likelihood slightly outperforms the robust method when both the percentage of background nodes is high and the linkage probability within community is low, in which case the robust method discards lots of information, while the multinomial pseudo-likelihood (or correspondingly degree corrected stochastic blockmodel) accounts for high variations on degrees. On the other hand, the robust method outperforms the multinomial pseudo-likelihood in all the other cases. In summary, the robust method has the best performance in terms of both accuracy and efficacy in almost all the setups we examined regardless the data follows stochastic blockmodels or not. In the only exception where the multinomial pseudo-likelihood method with logistic regression performs slightly better, the discrepancies between the two methods are small. Therefore, the robust community detection method is our recommended method.

7. Application

With the development of improved sequencing techniques, more and more de novo mutations in candidate genes associated with neurodevelopmental or neuropshychiatric diseases are being reported. Here we focus on autism spectrum disorder and related neurological disorders. Most identified de novo mutations are rare and patients with the same clinical symptoms often carry heterogeneous mutation loci on different genes. Most probably, the pathophysiology mechanism underpinning autism involves perturbed molecular pathways. There is evidence of enrichment of de novo mutations in gene groups connected by protein-protein interactions, co-expression patterns, or pathways defined by common functions, annotations or evolutional patterns (Allen et al., 2013). Our study targets interactive groups of biomarkers, *gene modules*, that form biological pathways producing autism. Gene modules are defined as a set of genes 1) whose product proteins interact on the molecular level and 2) expression levels change at the same time. Furthermore, we are particularly interested in autism related gene modules with 3) higher occurrences of de novo mutations in cases.

Autism and related disorder data from Hormozdiari et al. (2015) are employed, which reports four types of information (1. clinically diagnosed disease status, 2. RNA expression levels, 3. de novo mutations, 4. protein-protein interactions) from three major data consortiums including BrainSpan Atlas, published autism studies, protein-protein interaction databases. There are 52,801 verified protein-protein interaction links and 192,499 mRNA pairs with Pearson's correlation coefficient between their expression levels higher than 0·5, with an overlap of 1060 links. Together, there are 244,240 unique links from both data sources. These links involve 13,243 genes. Hormozdiari et al. (2015) further gathered the de novo mutation and length information on 796 out of the 13,243 genes. In total, 796 genes with de novo mutations are employed in our analysis with 1334 mutual links between them, among which 602 genes have at least one link and 194 have none. The number of gene groups are picked using a modified Bayesian information criterion designed specifically for stochastic blockmodels (Saldana et al., 2017), that is,

$$-2\mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}, \hat{P}; \hat{\boldsymbol{c}}, A) + \frac{(K+1)(K+2)}{2} \log\left(\frac{n(n-1)}{2}\right).$$

In this data of 796 genes, the model assuming seven autism related modules plus one irrelevant background group produces the smallest Bayesian information criterion.

Mutations are divided into two main categories – missense and loss of function. Synonymous mutations that differ at the DNA level but produce the same protein products are excluded. The frequencies of each type of mutation in a gene in all cases are summed up as well as the total number in the controls. Three covariates are employed in estimating the probability that a gene is involved in the occurrence or progression of autism and related neurological disorders – frequency of missense mutations in cases, frequency of loss of function mutations in cases, total number of mutations in controls. The choice of the covariates is based on biological beliefs on their involvement on autism development, hence decided a priori.

The robust community detection method in Section 4 identifies 53 genes showing no sign of involvement in autism or related disorders as well as 743 genes potentially involved in neurological disorder pathways. The 743 genes are clustered into seven non-overlapping gene modules with different sizes:7,533,79,28,35,43,18. The link densities,which are defined as the ratio of the number of links over the number of possible pairs, within each group and between any two groups are listed in Table 2. The majority of links concentrate on the diagonal of the linkage matrix for the seven groups related to autism. Group eight is the group of 53 irrelevant background genes, which have low to medium linkage probabilities with all groups including itself. The linkage probabilities for background genes within themselves are not necessarily higher than the linkage probabilities with autism-related genes in the other groups.

[Table 3 about here.]

The gene set enrichment analysis (GSEA) of the selected gene modules compared with the curated gene sets in the Molecular Signatures Database are listed in Table 3. P-values are calculated assuming a hypergeometric distribution for the number of overlapping genes between the selected group and the curated gene set. Given the large number of multiple comparisons, stringent P-value threshold 10^{-8} is employed. Group two overlaps significantly with ten gene sets in abnormal conditions such as carcinoma, cancer, UV response, apoptosis, Alzheimers and melanoma. Group three overlaps with gene sets related to neurological functions or disorders. Gene set “REACTOME AXON GUIDANCE” are genes involved in Axon guidance, the process by which neurons send out axons to reach the correct targets. Gene set “KEGG CALCIUM SIGNALING PATHWAY” concerns multiple cellular processes that uses calcium ions as the signal. Gene sets “REACTOME DEVELOPMENTAL BIOLOGY” and “REACTOME HEMOSTASIS” are composed of genes involved in developmental biology and hemostasis, respectively. Group four and seven overlaps with a lot cancer-related gene sets, while group one, five and six do not have any overlap with P-value less than 10^{-8}. Furthermore, our results are compared to those from the Merging Aaffected Genes into Integrated-Nnetworks method in Hormozdiari et al. (2015). The Merging Affected Genes into Integrated-Networks method was not able to detect group one. P-values from gene set enrichment analysis for the two best sets identified by their method against known neurodevelopmental diseases sets are $4{\cdot}2 \times 10^{-5}$ and $1{\cdot}0 \times 10^{-4}$, failing to reach the 10^{-8} threshold.

[Table 4 about here.]

8. Discussion

A major improvement of the proposed method over previous ones is the integration of network topology and auxiliary node information. The proposed analysis pools rich epigenomic information from heterogeneous online resources, such as expression/co-expression profiles from BrainSpan Atlas, de novo mutations in cases and controls from autism or related neurological disorder studies, protein-protein interactions in protein databases. Although these three types of information are measured on different cohorts, they describe distinct aspects of the candidate genes. They can be linked by unique genes, which are the unit of our analysis. In the era of big data, statistical methods need not be restricted to one data source or single clinical trial. Instead, methods should incorporate information from many related resources.

The estimation method is non-standard. For a fixed initial label assignment, we use the expectation-maximization algorithm to fit a pseudo-likelihood. Then the label assignment is updated according to the expectation-maximization results, and used as initial label assignment in the next iteration. Taking advantage of the pseudo-likelihood, we are able to allow heterogeneous linkage probabilities in the background. The consistency of the label assignments is proved for a simple version of this complicated procedure – one-step expectation-maximization. Further research is needed to understand the statistical properties of the algorithm in more complex settings.

Researchers have suggested that a node may belong to multiple communities in a biological networks. For example, Airoldi et al. (2008) proposed a mixed membership stochastic blockmodels and applied this model into a network of protein-protein interactions. We will explore the extension of the logistic regression augmented model to overlapping community detection in our future work.

Acknowledgements

Yunpeng Zhao was supported by the U.S. National Science Foundation Grant DMS 1513004. Qing Pan was supported by Award Numbers UL1TR000075 and KL2TR000076 from the NIH National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translational Sciences or the National Institutes of Health. The authors thank Professors J. G. Gastwirth for a critical reading of the original version of the paper.

Supplementary Materials

Web Appendix A, containing technical proofs of Theorem 1 and Lemma 1, is available with this paper at the Biometrics website on Wiley Online Library.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* **9,** 1981–2014.

Allen, A., Berkovic, S., Cossette, P., Delanty, N., Dlugos, D., Eichler, E., Epstein, M., Glauser, T., Goldstein, D., Han, Y., and et al. (2013). De novo mutations in epileptic encephalopathies. *Nature* **501,** 217–21.

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

Binkiewicz, N., Vogelstein, J. T., and Rohe, K. (2014). Covariate-assisted spectral clustering.

Choi, D. S., Wolfe, P. J., and Airoldi, E. M. (2012). Stochastic blockmodels with growing number of classes. *Biometrika* **99,** 273–284.

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

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.

Gormley, I. C. and Murphy, T. B. (2008). A mixture of experts model for rank data with applications in election studies. *The Annals of Applied Statistics* pages 1452–1477.

Gormley, I. C. and Murphy, T. B. (2010). A mixture of experts latent position cluster model for social network data. *Statistical methodology* **7,** 385–405.

Handcock, M. D., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *J. R. Statist. Soc. A* **170,** 301–354.

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

Hormozdiari, F., Penn, O., B., E., and Eichler, E. (2015). The discovery of integrated gene networks for autism and related disorders. *Genome Research* **9,** 1179–225.

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

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks* **31,** 204213.

Newman, M. and Clauset, A. (2016). Structure and inference in annotated networks. *Nature Communications* **7,**.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103,** 8577–8582.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures.

Journal of the American Statistical Association **96,** 1077–1087.

Saldana, D. F., Yu, Y., and Feng, Y. (2017). How many communities are there? *Journal of Computational and Graphical Statistics* **26,** 171–181.

Snijders, T. and Nowicki, K. (1997). Estimation and prediction for stochastic block-structures for graphs with latent block structure. *Journal of Classification* **14,** 75–100.

Tallberg, C. (2004). A bayesian approach to modeling stochastic blackstructures with covariates. *The Journal of Mathematical Sociology* **29,**.

Vinh, N. X., Epps, E., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research* **11,** 2837–54.

Xu, Z., Ke, Y., Wang, Y., Cheng, H., and Cheng, J. (2012). A model-based approach to attributed graph clustering. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD ’12, pages 505–516, New York, NY, USA. ACM.

Yan, B. and Sarkar, P. (2016). Convex relaxation for community detection with covariates.

Yang, J., McAuley, J., and Leskovec, J. (2013). Community Detection in Networks with Node Attributes. In *IEEE International Conference On Data Mining (ICDM)*.

Zhang, Y., Levina, E., and Zhu, J. (2015). Community detection in networks with node features.

Zhao, Y., Levina, E., and Zhu, J. (2011). Community extraction for social networks. *Proc. Nat. Acad. Sci.* **108,** 7321–7326.

Zhao, Y., Levina, E., and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Annals of Statistics* **40,** 2266–2292.

Received August 2016. *Revised February* 2008. *Accepted March* 2008.

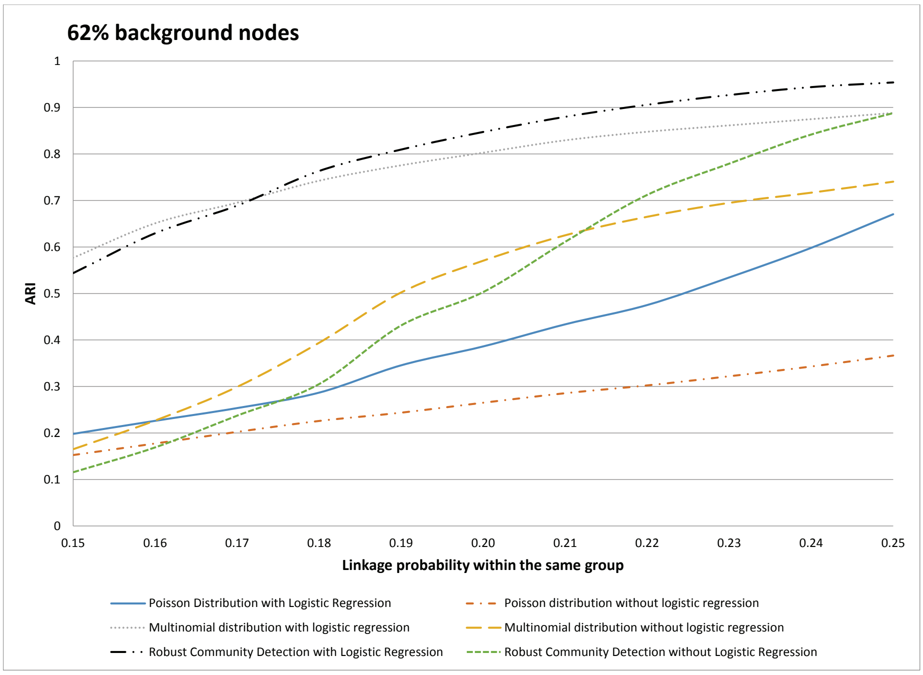

Figure 1. Comparison of the average ARI for Poisson pseudo-likelihood, multinomial pseudo-likelihood and robust community detection with and without logistic regressions under 68% of background nodes.

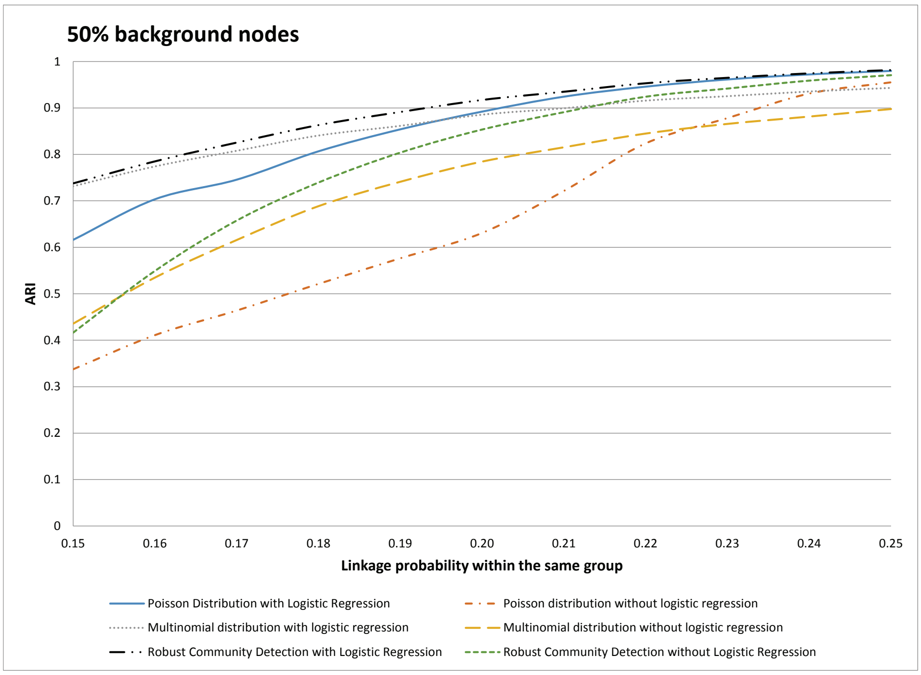

Figure 2. Comparison of the average ARI for Poisson pseudo-likelihood, multinomial pseudo-likelihood and robust community detection with and without logistic regressions under 50% of background nodes.

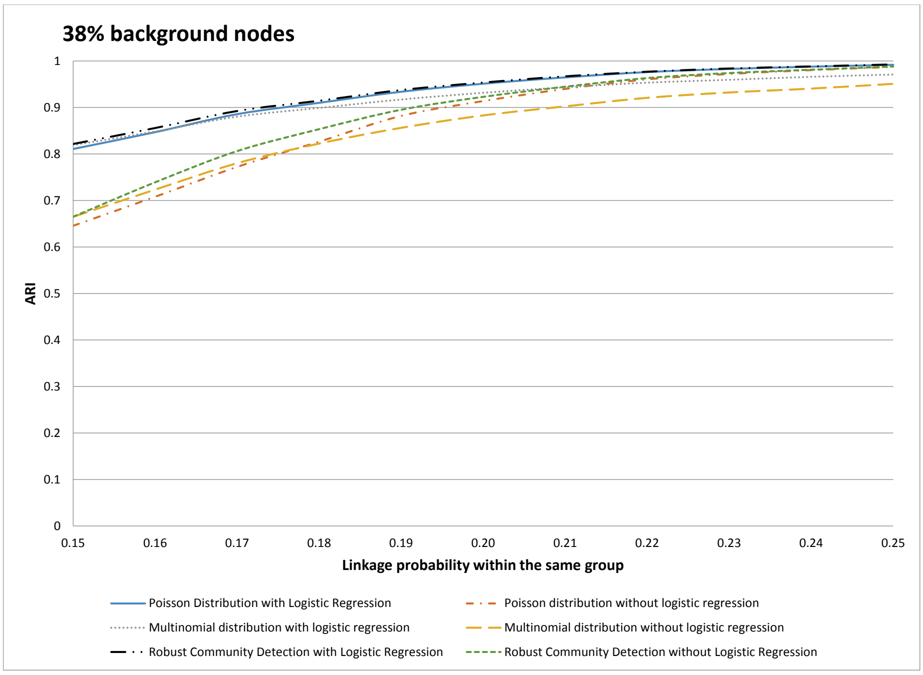

Figure 3. Comparison of the average ARI for Poisson pseudo-likelihood, multinomial pseudo-likelihood and robust community detection with and without logistic regressions under 32% of background nodes.

Table 1
Comparison of average adjusted rand index (ARI) $\times 100$ *under stochastic blockmodels. Numbers within parentheses are empirical standard deviations of ARI* $\times 100$.

	With Logistic Models			Without Logistic Models		
p_{11}	Poisson	Multinomial	Robust	Poisson	Multinomial	Robust
62% Background Nodels						
15	58 (12)	57 (13)	59 (12)	15 (7)	15 (8)	15 (8)
16	66 (8)	66 (9)	67 (8)	23 (11)	24 (11)	23 (11)
17	72 (7)	72 (6)	73 (7)	34 (13)	33 (13)	33 (13)
18	77 (5)	76 (5)	77 (5)	48 (14)	45 (13)	46 (15)
19	81 (5)	80 (5)	81 (4)	61 (11)	55 (13)	60 (11)
20	85 (4)	83 (4)	85 (4)	70 (8)	66 (9)	70 (9)
21	88 (3)	86 (4)	88 (3)	78 (6)	73 (7)	78 (7)
22	91 (3)	88 (3)	91 (3)	83 (4)	79 (6)	83 (5)
23	93 (3)	90 (3)	93 (3)	87 (4)	83 (5)	87 (4)
24	94 (2)	92 (3)	94 (2)	90 (3)	86 (4)	90 (3)
25	96 (2)	93 (2)	96 (2)	93 (3)	89 (3)	93 (3)
50% Background Nodels						
15	74 (5)	74 (5)	74 (5)	44 (10)	44 (10)	43 (11)
16	78 (4)	78 (4)	79 (4)	56 (8)	56 (8)	55 (10)
17	82 (4)	82 (4)	82 (4)	66 (6)	64 (7)	66 (7)
18	86 (3)	85 (4)	86 (3)	74 (6)	72 (6)	74 (6)
19	89 (3)	88 (3)	89 (3)	80 (5)	78 (5)	80 (5)
20	91 (3)	90 (3)	92 (3)	86 (4)	82 (4)	86 (4)
21	94 (2)	92 (3)	94 (2)	89 (3)	86 (4)	89 (3)
22	95 (2)	93 (2)	95 (2)	92 (3)	89 (3)	92 (3)
23	96 (2)	95 (2)	97 (2)	94 (2)	91 (3)	94 (2)
24	98 (1)	96 (2)	97 (1)	96 (2)	93 (3)	96 (2)
25	98 (1)	96 (2)	98 (1)	97 (1)	94 (2)	97 (1)
38% Background Nodels						
15	82 (4)	82 (4)	82 (3)	67 (6)	67 (6)	67 (6)
16	86 (3)	86 (3)	86 (3)	74 (5)	74 (5)	74 (5)
17	89 (3)	88 (3)	89 (3)	81 (4)	80 (4)	80 (4)
18	91 (3)	91 (3)	91 (3)	85 (3)	84 (4)	85 (3)
19	94 (2)	92 (2)	94 (2)	89 (3)	87 (3)	89 (3)
20	96 (2)	94 (2)	96 (2)	92 (2)	90 (3)	92 (2)
21	97 (1)	95 (2)	97 (1)	95 (2)	92 (2)	95 (2)
22	98 (1)	96 (2)	98 (1)	96 (2)	94 (2)	96 (2)
23	98 (1)	97 (1)	98 (1)	97 (1)	95 (2)	97 (1)
24	99 (1)	97 (1)	99 (1)	98 (1)	96 (2)	98 (1)
25	99 (1)	98 (1)	99 (1)	99 (1)	97 (2)	99 (1)

Table 2
Comparison of average adjusted rand index (ARI) $\times 100$ under heterogeneous backgrounds. Numbers within parentheses are empirical standard deviations of ARI $\times 100$.

	With Logistic Models			Without Logistic Models		
p_{11}	Poisson	Multinomial	Robust	Poisson	Multinomial	Robust
62% Background Nodels						
15	20 (11)	58 (14)	54 (21)	15 (6)	17 (8)	12 (10)
16	23 (13)	65 (9)	63 (18)	18 (5)	23 (10)	17 (12)
17	25 (13)	70 (8)	69 (16)	20 (5)	30 (12)	24 (17)
18	29 (14)	74 (6)	76 (11)	23 (5)	39 (13)	31 (21)
19	35 (18)	78 (5)	81 (8)	24 (5)	50 (12)	43 (26)
20	39 (20)	80 (5)	85 (6)	27 (6)	57 (12)	50 (27)
21	43 (23)	83 (5)	88 (5)	29 (5)	63 (10)	61 (27)
22	48 (25)	85 (4)	91 (3)	30 (6)	66 (10)	71 (25)
23	53 (27)	86 (4)	93 (3)	32 (7)	69 (10)	78 (22)
24	60 (29)	87 (4)	94 (2)	34 (9)	72 (10)	84 (19)
25	67 (30)	89 (4)	95 (2)	37 (13)	74 (10)	89 (15)
50% Background Nodels						
15	62 (19)	73 (5)	74 (5)	34 (12)	44 (9)	42 (12)
16	70 (15)	77 (4)	79 (4)	41 (15)	53 (9)	55 (11)
17	75 (14)	81 (4)	83 (4)	46 (15)	62 (8)	66 (8)
18	81 (10)	84 (4)	86 (3)	52 (15)	69 (6)	74 (7)
19	85 (9)	86 (4)	89 (3)	58 (15)	74 (6)	80 (6)
20	89 (7)	89 (3)	92 (3)	63 (17)	78 (5)	85 (5)
21	92 (4)	90 (3)	93 (2)	72 (18)	82 (5)	89 (3)
22	95 (2)	92 (3)	95 (2)	82 (16)	84 (5)	92 (2)
23	96 (2)	93 (2)	96 (2)	88 (15)	87 (4)	94 (2)
24	97 (2)	94 (2)	97 (1)	93 (10)	88 (4)	96 (2)
25	98 (1)	94 (2)	98 (1)	96 (7)	90 (4)	97 (1)
38% Background Nodels						
15	81 (5)	82 (4)	82 (4)	65 (8)	66 (6)	67 (7)
16	85 (4)	85 (3)	86 (3)	71 (7)	72 (5)	74 (5)
17	89 (3)	88 (3)	89 (3)	77 (7)	78 (5)	81 (4)
18	91 (3)	90 (3)	91 (2)	83 (6)	82 (4)	85 (4)
19	93 (2)	92 (3)	94 (2)	88 (4)	86 (4)	90 (3)
20	95 (2)	93 (2)	95 (2)	91 (3)	88 (3)	92 (3)
21	96 (2)	94 (2)	97 (2)	94 (2)	90 (3)	94 (2)
22	98 (1)	95 (2)	98 (1)	96 (2)	92 (3)	96 (2)
23	98 (1)	96 (2)	98 (1)	97 (2)	93 (2)	97 (1)
24	99 (1)	97 (2)	99 (1)	98 (1)	94 (2)	98 (1)
25	99 (1)	97 (1)	99 (1)	99 (1)	95 (2)	99 (1)

Table 3
Estimated Link Probabilities$\times 10^3$ *between Groups*

Group 1–7							*Group 8*
204	0	0	0	0	0	0	*3*
0	**0**	1	1	0	0	1	*2*
0	1	**23**	0	11	3	1	*16*
0	1	0	**628**	221	0	4	*86*
0	0	11	221	**175**	0	0	*65*
0	0	3	0	0	**30**	0	*2*
0	1	1	4	0	0	**414**	*14*
3	*2*	*16*	*86*	*65*	*2*	*14*	*53*

Table 4

Gene Set Enrichment Analysis of Selected Groups

Group Number	Gene Set Name	Group Size	GeneSet Size	Overlap Size	Nominal P-value	FDR q-value
2	DODD NASOPHARYNGEAL CARCINOMA UP	533	1821	62	$2{\cdot}36 \times 10^{-14}$	$7{\cdot}22 \times 10^{-11}$
2	GOBERT OLIGODENDROCYTE DIFFERENTIATION DN	533	1080	46	$3{\cdot}05 \times 10^{-14}$	$7{\cdot}22 \times 10^{-11}$
2	BONOME OVARIAN CANCER SURVIVAL SUBOPTIMAL	533	510	31	$6{\cdot}51 \times 10^{-14}$	$1{\cdot}03 \times 10^{-10}$
2	NABA MATRISOME	533	1028	42	$1{\cdot}62 \times 10^{-12}$	$1{\cdot}92 \times 10^{-9}$
2	SENESE HDAC3 TARGETS DN	533	536	29	$7{\cdot}37 \times 10^{-12}$	$6{\cdot}97 \times 10^{-8}$
2	YOSHIMURA MAPK8 TARGETS UP	533	1305	46	$2{\cdot}09 \times 10^{-11}$	$1{\cdot}65 \times 10^{-8}$
2	DACOSTA UV RESPONSE VIA ERCC3 DN	533	855	36	$2{\cdot}91 \times 10^{-11}$	$1{\cdot}93 \times 10^{-8}$
2	GRAESSMANN APOPTOSIS BY DOXORUBICIN DN	533	1781	55	$3{\cdot}27 \times 10^{-11}$	$1{\cdot}93 \times 10^{-8}$
2	BLALOCK ALZHEIMERS DISEASE UP	533	1691	53	$4{\cdot}49 \times 10^{-11}$	$2{\cdot}19 \times 10^{-8}$
2	ONKEN UVEAL MELANOMA UP	533	783	34	$4{\cdot}64 \times 10^{-11}$	$2{\cdot}19 \times 10^{-8}$
3	REACTOME AXON GUIDANCE	79	251	9	$5{\cdot}53 \times 10^{-10}$	$2{\cdot}11 \times 10^{-6}$
3	KEGG CALCIUM SIGNALING PATHWAY	79	178	8	$8{\cdot}92 \times 10^{-10}$	$2{\cdot}11 \times 10^{-6}$
3	REACTOME DEVELOPMENTAL BIOLOGY	79	396	10	$1{\cdot}71 \times 10^{-9}$	$2{\cdot}70 \times 10^{-6}$
3	REACTOME HEMOSTASIS	79	466	10	$8{\cdot}06 \times 10^{-9}$	$9{\cdot}52 \times 10^{-6}$
4	DACOSTA UV RESPONSE VIA ERCC3 DN	28	855	11	$1{\cdot}40 \times 10^{-12}$	$6{\cdot}60 \times 10^{-9}$
4	GRAESSMANN APOPTOSIS BY DOXORUBICIN DN	28	1781	13	$9{\cdot}26 \times 10^{-12}$	$2{\cdot}19 \times 10^{-8}$
4	MILI PSEUDOPODIA HAPTOTAXIS UP	28	518	8	$6{\cdot}30 \times 10^{-10}$	$9{\cdot}92 \times 10^{-7}$
7	GOBERT OLIGODENDROCYTE DIFFERENTIATION UP	18	570	11	$2{\cdot}86 \times 10^{-17}$	$1{\cdot}35 \times 10^{-13}$
7	DUTERTRE ESTRADIOL RESPONSE 24HR UP	18	324	9	$1{\cdot}77 \times 10^{-15}$	$4{\cdot}19 \times 10^{-12}$
7	PUJANA BRCA2 PCC NETWORK	18	423	9	$1{\cdot}97 \times 10^{-14}$	$3{\cdot}10 \times 10^{-11}$
7	PUJANA XPRSS INT NETWORK	18	168	7	$2{\cdot}37 \times 10^{-13}$	$2{\cdot}63 \times 10^{-10}$
7	GEORGES TARGETS OF MIR192 AND MIR215	18	893	10	$2{\cdot}78 \times 10^{-13}$	$2{\cdot}63 \times 10^{-10}$
7	NUYTTEN EZH2 TARGETS DN	18	1024	10	$1{\cdot}08 \times 10^{-12}$	$8{\cdot}48 \times 10^{-10}$
7	PUJANA CHEK2 PCC NETWORK	18	779	9	$4{\cdot}68 \times 10^{-12}$	$3{\cdot}16 \times 10^{-9}$
7	KINSEY TARGETS OF EWSR1 FLII FUSION UP	18	1278	9	$3{\cdot}75 \times 10^{-10}$	$2{\cdot}22 \times 10^{-7}$
7	PUJANA BRCA CENTERED NETWORK	18	117	5	$8{\cdot}19 \times 10^{-10}$	$4{\cdot}30 \times 10^{-7}$
7	BLUM RESPONSE TO SALIRASIB DN	18	342	6	$2{\cdot}80 \times 10^{-9}$	$1{\cdot}18 \times 10^{-6}$

[1]

The first column is the group number identified by the proposed method; Size refers to the number of genes in the identified group, or gene set in the GSEA or their overlap.