2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834456

Exact Recovery Threshold in Dynamic Binary Censored Block Model

Javad Zahedi Moghaddam, Mohammad Esmaeili, and Aria Nosratinia
Department of Electrical and Computer Engineering
The University of Texas at Dallas
Richardson, TX 75083-0688, USA
{moghaddam.jz, esmaeili, aria}@utdallas.edu

***Abstract*—This paper considers community detection in the dynamic binary censored block model. Under this model, the graph is observed at successive times (snapshots), and the node label in the current snapshot is dependent on the same node label in the previous τ snapshots. In this paper, the maximum likelihood estimator of the current node labels is obtained under this model, subject to the observation of the graph in the present and past snapshots, and the exact recovery conditions are derived. Relaxing the maximum likelihood estimator, a semidefinite programming algorithm is proposed for community detection. In the asymptotic regime, a sufficient condition for exact recovery is obtained using the semidefinite programming estimator, which is shown to asymptotically match the sufficient conditions for exact recovery.**

***Index Terms*—Community detection, dynamic, censored block model, semidefinite programming, Markov models**

I. Introduction

Community detection is an unsupervised learning category in network science and signal processing [1]–[3] whose applications include discovery of hidden relationships among individuals in social networks [4], identification of protein complexes in biology [5], and transportation networks [6].

The asymptotic performance of community detection algorithms has been studied under models such as stochastic block model (SBM), censored block model (CBM), and latent space model. These models provide intuitions for the analysis of real-world networks. Recently, heightened attention has been taken towards the SBM and CBM [7]–[13] as generative models, where the graph edges are drawn independently conditioned on the nodes' labels. Though the maximum likelihood estimator is the optimum estimator for community detection under the SBM and CBM in the dense regime, it is NP-hard and intractable practically [14]. To address this issue, some alternative methods have been exploited in the literature including spectral clustering [8], [15], semidefinite programming (SDP) [7], [11], [12], [16], and belief propagation [10].

Most of the research works still have focused on static community detection, while many real-world networks are dynamic in nature, and the community membership of nodes and edges of a graph evolve through time. Even though some research [17]–[21] in dynamic SBM have conducted to find certain regimes for community detection, extracting tight bounds for dynamic community detection has remained an important trend of study, especially for the dynamic CBM. This paper considers extracting certain regimes in which the exact recovery of the dynamic binary CBM has a solution. To this end, we first assume that there exists only a stochastic model for node labels between τ sequential snapshots while edges are independent. Then, under the dynamic binary CBM, the achievability bound of the exact recovery is calculated using the semidefinite programming algorithm. In addition, the converse is proven by finding regimes where the maximum likelihood estimator fails. We will show that achievability and converse bounds are tight asymptotically.[1]

In the remainder of this paper, Section II explains model and settings, and Section III explores dynamic community detection when there is a first-order Markov dynamic dependency between node labels. Section IV assumes the current label is dependent on a fixed sequence of its previous labels, i.e. $\tau \geq 2$. It is shown the solution is similar to $\tau = 1$, and we only need to update some scalars and vectors with respect to the new value of τ. Finally, Section VI concludes the paper.

II. Model and Settings

The standard CBM consists of an *Erdős–Rényi* graph $\boldsymbol{G}$ with n nodes, node labels $g_i \in \{-1, +1\}$ and edge labels $E_{ij} \in \{+1, -1, 0\}$ are drawn according to the following probability density

$$E_{ij} \sim \begin{cases} p(1-\xi)\delta_{+1} + p\xi\delta_{-1} + (1-p)\delta_0 & \text{if } g_i = g_j \\ p(1-\xi)\delta_{-1} + p\xi\delta_{+1} + (1-p)\delta_0 & \text{if } g_i \neq g_j \end{cases} \tag{1}$$

where $p = a\frac{\ln n}{n}$ for a fixed $a > 0$, $\xi \in [0, 1/2]$ and δ is the Dirac delta function centered at its subscript. The log-likelihood in the CBM is known to be [11]

$$\ln \mathbb{P}(\boldsymbol{A}|\boldsymbol{g}) = \frac{1}{4}\ln\left(\frac{1-\xi}{\xi}\right)\boldsymbol{g}^T\boldsymbol{A}\boldsymbol{g} + C_0, \tag{2}$$

This work was supported in part by the grant 2008684 from the National Science Foundation.

[1]A different variation in the graph model was analyzed by the notion of side information [10], [11], [22]. Side information consists of a variable for each node that, given the label of that node, is independent of the graph and all other variables. Side information models extraneous information that is useful to graph inference but does not possess a graph structure. This paper proposes and analyzes a model with a different statistical structure that is aimed at understanding and analyzing community detection under temporal dynamics.

where $\boldsymbol{A}$ is the adjacency matrix and $\boldsymbol{g}$ is the vector of node labels, and C_0 is a deterministic scalar. Our model involves multiple temporal snapshots of the graph $\boldsymbol{G}(t)$ at times t, with adjacency matrices $\boldsymbol{A}(t)$ and node label vectors $\boldsymbol{g}(t)$. In the dynamic binary CBM, the log likelihood is expressed as follows

$$\ln \mathbb{P}\big(\boldsymbol{A}(t), \boldsymbol{g}(t) | \{\boldsymbol{g}(u)\}_{t-\tau}^{t-1}\big) = \ln \mathbb{P}\big(\boldsymbol{A}(t) | \boldsymbol{g}(t)\big) + Q, \quad (3)$$

such that

$$\begin{aligned} Q &\triangleq \ln \mathbb{P}\big(\boldsymbol{g}(t) | \{\boldsymbol{g}(u)\}_{t-\tau}^{t-1}\big) \\ &= \sum_{i=1}^{n} \ln \mathbb{P}\big(g_i(t) | \{g_i(u)\}_{t-\tau}^{t-1}\big), \end{aligned}$$

where the last equality holds since node labels are independent and identically distributed.

III. Analysis of First-Order Markov Dependency for Labels

We begin by showing the main results of the paper for first-order Markov labels. Subsequently, we generalize the results to any finite-memory temporal dependency for labels.

Calculating and substituting the term Q in (3) using first-order Markov relations for label probabilities, we obtain

$$\begin{aligned} \ln \mathbb{P}\big(\boldsymbol{A}(t), \boldsymbol{g}(t) | \boldsymbol{g}(t-1)\big) &= C_1\, \boldsymbol{g}^T(t) \boldsymbol{A}(t) \boldsymbol{g}(t) \\ &\quad + C_2\, \boldsymbol{g}^T(t) \boldsymbol{g}(t-1) + C_3, \quad (4) \end{aligned}$$

where $C_1 = 0.25 \ln\big(\frac{1-\xi}{\xi}\big)$, $C_2 = 0.5 \ln \frac{\eta}{1-\eta}$, and η is the usual Markov parameter indicating the probability of persistence in a binary state. In (4), C_3 is a constant which will be omitted due to its insignificance in the next analyses. Note that, C_2 is either a constant or a function of graph size n, while $\eta \in [0, 1]$.

Let $\boldsymbol{Z}(t) \triangleq \boldsymbol{g}(t)\boldsymbol{g}^T(t)$ and $\boldsymbol{Z}^*(t) \triangleq \boldsymbol{g}^*(t)\boldsymbol{g}^{*T}(t)$ where $\boldsymbol{g}^*(t)$ shows the true label vector at t. The maximum likelihood estimator of (4) is formulated as

$$\begin{aligned} \hat{\boldsymbol{g}}(t) = \underset{\boldsymbol{g}(t)}{\arg\max} \quad & C_1 \big\langle \boldsymbol{Z}(t), \boldsymbol{A}(t) \big\rangle + C_2 \big\langle \boldsymbol{g}(t), \boldsymbol{g}(t-1) \big\rangle \\ \text{subject to} \quad & rank\big(\boldsymbol{Z}(t)\big) = 1 \\ & Z_{ii}(t) = 1, \;\; i \in [n] \end{aligned} \quad (5)$$

where $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = trace\left(\boldsymbol{Y}^T \boldsymbol{X}\right)$. Define

$$f\left(\boldsymbol{g}(t)\right) \triangleq C_1 \big\langle \boldsymbol{Z}(t), \boldsymbol{A}(t) \big\rangle + C_2 \big\langle \boldsymbol{g}(t), \boldsymbol{g}(t-1) \big\rangle.$$

Use the Schur's Lemma to relax the rank one constraint in (5) with $\boldsymbol{Z}(t) - \boldsymbol{g}(t)\boldsymbol{g}^T(t) \succeq 0$ results in

$$\begin{aligned} \hat{\boldsymbol{g}}(t) = \underset{\boldsymbol{g}(t)}{\arg\max} \quad & f(\boldsymbol{g}(t)) \\ \text{subject to} \quad & \boldsymbol{R} = \begin{bmatrix} 1 & \boldsymbol{g}^T(t) \\ \boldsymbol{g}(t) & \boldsymbol{Z}(t) \end{bmatrix} \succeq 0, \\ & Z_{ii}(t) = 1, \;\; i \in [n]. \end{aligned} \quad (6)$$

Define $\alpha \triangleq \lim_{n\to\infty} \frac{C_2}{2\ln n}$, and $\beta \triangleq \sqrt{\alpha^2 + 4a^2 C_1^2 \xi(1-\xi)}$, and

$$\mu(a, \alpha) \triangleq a - \frac{\beta}{C_1} + \frac{\alpha}{2C_1} \ln\left(\frac{(1-\xi)(\alpha+\beta)}{\xi(\beta-\alpha)}\right).$$

The achievability and converse for exact recovery of the dynamic binary CBM are represented in the following theorems:

Theorem 1. *The semidefinite programming estimator of* (6) *is asymptotically optimal, if*

$$\begin{cases} \mu(a,\alpha) > 1 & \textit{when} \;\; 0 \le \alpha < a\, C_1(1-2\xi) \\ \alpha > 1 & \textit{when} \;\; \alpha \ge a\, C_1(1-2\xi) \end{cases}.$$

Proof. See Section V-A. □

Theorem 2. *Under the dynamic binary CBM, for any sequence of estimators* $\hat{\boldsymbol{Z}}_n(t)$, $\mathbb{P}\big(\hat{\boldsymbol{Z}}_n(t) = \boldsymbol{Z}^*(t)\big) \to 0$ *as* $n \to \infty$, *if*

$$\begin{cases} \mu(a,\alpha) < 1 & \textit{when} \;\; 0 \le \alpha < a\, C_1(1-2\xi) \\ \alpha < 1 & \textit{when} \;\; \alpha \ge a\, C_1(1-2\xi) \end{cases}.$$

Proof. See Section V-B. □

IV. Generalization to Arbitrary Temporal Dependency for Labels

In this section, we generalize the temporal dependency such that the current label of a node is dependent on a sequence of its previous labels, i.e. $\tau \ge 2$. This means

$$\mathbb{P}\big(g_i(t) | \{g_i(u)\}_{t-\tau}^{t-1}\big) = \prod_{j=1}^{2^\tau} \theta_j^{m_j^i}, \quad (7)$$

where $0 \le \theta_j \le 1$, $\tau \in \{0, 1, \ldots, t-1\}$. Since we consider τ previous snapshots and our clusters are binary, the sequence of the previous labels has 2^τ different forms. In (7), m_j^i is the j-th sequence among 2^τ candidates. For example, m_1^i indicates that the previous τ labels are similar to the current label of node i; m_2^i shows that the previous $\tau - 1$ labels are similar to the current label, and $g_i(t-\tau) = -g_i(t)$; and by continuing this procedure for all the possible sequence of labels, $m_{2^\tau}^i$ implies that the previous τ labels of node i are different from its current label. In other words,

$$\begin{aligned} m_1^i &= \prod_{k=1}^{\tau} \mathbf{1}(g_i(t) = g_i(t-k)), \\ m_2^i &= \mathbf{1}(g_i(t) \ne g_i(t-\tau)) \prod_{k=1}^{\tau-1} \mathbf{1}(g_i(t) = g_i(t-k)), \\ &\vdots \\ m_{2^\tau}^i &= \prod_{k=1}^{\tau} \mathbf{1}(g_i(t) \ne g_i(t-k)), \end{aligned}$$

where $\mathbf{1}(\cdot)$ is the indicator function. Hence, Q in (3) can be rewritten as

$$\begin{aligned} Q &= \sum_{i=1}^{n} \Big[\sum_{j=1}^{2^\tau} \ln(\theta_j) m_j^i \Big] \\ &= \sum_{j=1}^{2^\tau} \ln(\theta_j) \Big[\sum_{i=1}^{n} m_j^i \Big]. \end{aligned} \quad (8)$$

Since we work on binary mode $g_i(t) \in \{-1,+1\}$,

$$\begin{cases} \mathbf{1}(g_i(t) = g_i(u)) = \frac{1+g_i(t)g_i(u)}{2} & \forall i,u,t \\ [g_i(u)]^{2r} = 1 & \forall i,u,r \\ [g_i(u)]^{2r-1} = g_i(u) & \forall i,u,r. \end{cases}$$

Therefore,

$$Q = T_0(\boldsymbol{\theta})H\big(\{\boldsymbol{g}(u)\}_{t-\tau}^{t-1},\boldsymbol{\theta}\big) + T_1(\boldsymbol{\theta})\boldsymbol{g}^T(t)\boldsymbol{F}\big(\{\boldsymbol{g}(u)\}_{t-\tau}^{t-1},\boldsymbol{\theta}\big), \tag{9}$$

where $\boldsymbol{\theta} = [\theta_1,\theta_2,\ldots,\theta_{2^\tau}]^T$, H is a real-valued function, and $\boldsymbol{F}$ is a vector-valued function of $\{\boldsymbol{g}(u)\}_{t-\tau}^{t-1}$ and $\boldsymbol{\theta}$. In (9), T_0 and T_1 are real-valued functions of $\boldsymbol{\theta}$. Let $w_1 = C_1 = \ln\big(\frac{1-\xi}{\xi}\big)$, $w_2 = T_1(\boldsymbol{\theta})$, $\boldsymbol{Z}(t) \triangleq \boldsymbol{g}(t)\boldsymbol{g}^T(t)$, and

$$\mathcal{P} \triangleq w_1 \left\langle \boldsymbol{Z}(t),\boldsymbol{A}(t)\right\rangle + w_2 \left\langle \boldsymbol{g}(t),\boldsymbol{F}\big(\{\boldsymbol{g}(u)\}_{u=t-\tau}^{t-1}\big)\right\rangle. \tag{10}$$

By inserting (9) and (2) in (3) and considering (10), the community detection problem will be equal to the maximum likelihood estimator as follows

$$\begin{aligned} \hat{\boldsymbol{g}}(t) = \arg\max_{\boldsymbol{g}(t)} \quad & \mathcal{P} \\ \text{subject to} \quad & rank\big(\boldsymbol{Z}(t)\big) = 1, \\ & Z_{ii}(t) = 1, \;\; i \in [n]. \end{aligned} \tag{11}$$

Equation (11) shows solving the dynamic community detection problem in a general case is similar to the first-order Markov case in Section III. Indeed, $\boldsymbol{F}\big(\{\boldsymbol{g}(u)\}_{t-\tau}^{t-1}\big) = \boldsymbol{g}(t-1)$ and $w_2 = C_2$ for the first-order Markov dependency. Therefore, one just needs to calculate $\boldsymbol{F}\big(\{\boldsymbol{g}(u)\}_{t-\tau}^{t-1}\big)$ and w_2 for an arbitrary fixed τ, and then (11) can be formulated and be solved by following the instructions applied on (5) to address the dynamic community detection in binary CBMs. In the following example, the problem is solved for the dynamic binary CBM, where $\tau = 2$.

A. Example: semidefinite programming for dynamic binary CBM where $\tau = 2$

For the sake of brevity, here

$$Q = \sum_{j=1}^{4} \ln(\theta_j)\big[\sum_{i=1}^{n} m_j^i\big], \tag{12}$$

where

$$\begin{aligned} m_1^i &= \mathbf{1}(g_i(t) = g_i(t-1))\mathbf{1}(g_i(t) = g_i(t-2)), \\ m_2^i &= \mathbf{1}(g_i(t) = g_i(t-1))\mathbf{1}(g_i(t) \neq g_i(t-2)), \\ m_3^i &= \mathbf{1}(g_i(t) \neq g_i(t-1))\mathbf{1}(g_i(t) = g_i(t-2)), \\ m_4^i &= \mathbf{1}(g_i(t) \neq g_i(t-1))\mathbf{1}(g_i(t) \neq g_i(t-2)), \end{aligned}$$

Now, by expanding Q in (12) and using the properties of binary community detection,

$$\begin{aligned} &\boldsymbol{F}\big(\{\boldsymbol{g}(u)\}_{t-2}^{t-1}\big) = \boldsymbol{g}(t-1) - \big[\ln^2\frac{\theta_1}{\theta_4} - \ln^2\frac{\theta_2}{\theta_3}\big]\boldsymbol{g}(t-2), \\ &w_2 = \frac{1}{4}\ln\big(\frac{\theta_1\theta_2}{\theta_3\theta_4}\big). \end{aligned} \tag{13}$$

Applying (13) in (11), the achievability and converse for exact recovery of the dynamic binary CBM where $\tau = 2$ will obtained in the following theorems.

Let $\boldsymbol{Z}^*(t) \triangleq \boldsymbol{g}^*(t)\boldsymbol{g}^{*T}(t)$ where $\boldsymbol{g}^{*T}(t)$ shows the true label vector at t. Define $\omega \triangleq \lim_{n\to\infty}\frac{w_2}{2\ln n}$, $\phi \triangleq \sqrt{\omega^2 + 4a^2 w_1^2 \xi(1-\xi)}$, and

$$\psi(a,\omega) \triangleq a - \frac{\phi}{w_1} + \frac{\omega}{2w_1}\ln\big(\frac{(1-\xi)(\omega+\phi)}{\xi(\phi-\omega)}\big).$$

Theorem 3. *The semidefinite programming estimator of* (11) *is asymptotically optimal, if*

$$\begin{cases} \psi(a,\omega) > 1 & \textit{when} \;\; 0 \le \omega < a\,w_1(1-2\xi) \\ \omega > 1 & \textit{when} \;\; \textit{if}\, \omega \ge a\,w_1(1-2\xi) \end{cases}.$$

Proof. Similar to the proof of Theorem 1. □

Theorem 4. *For any sequence of estimators $\hat{\boldsymbol{Z}}_n(t)$, $\mathbb{P}\big(\hat{\boldsymbol{Z}}_n(t) = \boldsymbol{Z}^*(\boldsymbol{t})\big) \to 0$ as $n \to \infty$, if*

$$\begin{cases} \psi(a,\omega) < 1 & \textit{when} \;\; 0 \le \omega < a\,w_1(1-2\xi) \\ \omega < 1 & \textit{when} \;\; \textit{if}\, \omega \ge a\,w_1(1-2\xi) \end{cases}.$$

Proof. Similar to the proof of Theorem 2. □

It should be noted that Theorems 3 and 4 provide the strict bounds for general case of dynamic community detection in CBMs when $\tau = 2$.

V. PROOFS

A. Proof of Theorem 1

Lemma 1. *Consider the diagonal matrix $\boldsymbol{D}^*$ and symmetric $\boldsymbol{S}^* \succeq 0$ such that the second* smallest *eigenvalue $\lambda_2(\boldsymbol{S}^*) > 0$, with the following submatrices*

$$\boldsymbol{S}^* = \begin{bmatrix} \boldsymbol{S}_1^* & \boldsymbol{S}_2^{*T} \\ \boldsymbol{S}_2^* & \boldsymbol{S}_3^* \end{bmatrix},$$

where

$$\begin{cases} \boldsymbol{S}_1^* = 0.5\,C_2\,\boldsymbol{g}^T(t-1)\boldsymbol{g}^*(t) \\ \boldsymbol{S}_2^* = -0.5\,C_2\,\boldsymbol{g}(t-1) \\ \boldsymbol{S}_3^* = \boldsymbol{D}^* - C_1\boldsymbol{A}(t) \\ \boldsymbol{S}^{*T}[1,\boldsymbol{g}^{*T}(t)]^T = \boldsymbol{0}. \end{cases}$$

Then, $(\boldsymbol{D}^*,\boldsymbol{S}^*)$ is the dual optimal solution of (6)*, whose unique primal solution is $\hat{\boldsymbol{Z}}(t) = \boldsymbol{g}^*(t)\boldsymbol{g}^{*T}(t)$.*

Proof. The Lagrangian of (6) is given by

$$\begin{aligned} L(\boldsymbol{Z}(t),\boldsymbol{g}(t),\boldsymbol{S},\boldsymbol{D}) =& f(\boldsymbol{Z}(t),\boldsymbol{g}(t)) \\ &+ \big\langle \boldsymbol{S},\boldsymbol{R}\big\rangle - \big\langle \boldsymbol{D},\boldsymbol{Z}(t) - \mathbf{I}\big\rangle, \end{aligned}$$

where $\boldsymbol{S} \succeq 0$ and $\boldsymbol{D} = \mathrm{diag}(d_i)$ are Lagrange multipliers. It can be proven for any $\boldsymbol{Z}(t)$ satisfying optimization in (6) that

$$f(\boldsymbol{Z}(t),\boldsymbol{g}(t)) \le L(\boldsymbol{Z}(t),\boldsymbol{g}(t),\boldsymbol{S}^*,\boldsymbol{D}^*) = f(\boldsymbol{Z}^*(t),\boldsymbol{g}^*(t)),$$

where it holds since $\big\langle \boldsymbol{S}^*,\boldsymbol{R}\big\rangle \ge 0$, $Z_{ii}(t) = 1$ for all $i \in [n]$, $\boldsymbol{S}_1^* = -\boldsymbol{S}_2^{*T}\boldsymbol{g}^*(t)$, and $\boldsymbol{S}_2^* = -\boldsymbol{S}_3^*\boldsymbol{g}^*(t)$. Hence, $\boldsymbol{Z}^*(t) =$

$\boldsymbol{g}^*(t)\boldsymbol{g}^{*T}(t)$ is a primal optimal solution. Assuming $\tilde{\boldsymbol{Z}}(t)$ is another optimal solution, thus

$$\begin{aligned}&\langle \boldsymbol{S}^*, \tilde{\boldsymbol{R}}\rangle = \boldsymbol{S}_1^* + 2\langle \boldsymbol{S}_2^*, \tilde{\boldsymbol{g}}(t)\rangle + \langle \boldsymbol{D}^* - C_1\boldsymbol{A}(t), \tilde{\boldsymbol{Z}}(t)\rangle\\ &\overset{(a)}{=} \boldsymbol{S}_1^* + 2\langle \boldsymbol{S}_2^*, \boldsymbol{g}^*(t)\rangle + \langle \boldsymbol{D}^*, \boldsymbol{Z}^*(t)\rangle - C_1\langle \boldsymbol{A}(t), \boldsymbol{Z}^*(t)\rangle\\ &= \langle \boldsymbol{S}^*, \boldsymbol{R}^*\rangle = 0,\end{aligned}$$

where (a) holds because $\langle \boldsymbol{A}(t), \boldsymbol{Z}^*(t)\rangle = \langle \boldsymbol{A}(t), \tilde{\boldsymbol{Z}}(t)\rangle$, $Z_{ii}^*(t) = \tilde{Z}_{ii}(t) = 1$ for all $i \in [n]$, and $\langle \boldsymbol{S}_2^*, \tilde{\boldsymbol{g}}(t)\rangle = \langle \boldsymbol{S}_2^*, \boldsymbol{g}^*(t)\rangle$. Since $\tilde{\boldsymbol{R}} \succeq 0$ and $\boldsymbol{S}^* \succeq 0$ while $\lambda_2(\boldsymbol{S}^*)$ is positive, $\tilde{\boldsymbol{R}}$ must be a multiple of $\boldsymbol{R}^*$. Also, since $Z_{ii}^*(t) = \tilde{Z}_{ii}(t) = 1$ for all $i \in [n]$, we have $\tilde{\boldsymbol{R}} = \boldsymbol{R}^*$. □

In addition, $\boldsymbol{S}^*$ must satisfy conditions in Lemma 1 with probability $1 - o(1)$. Let

$$d_i^* = C_1 \sum_{j=1}^{n} A_{ij}(t)\, g_j^*(t)\, g_i^*(t) + 0.5\, C_2 g_i(t-1) g_i^*(t). \quad (14)$$

Then,

$$\boldsymbol{D}^*\boldsymbol{g}^*(t) = C_1\boldsymbol{A}(t)\boldsymbol{g}^*(t) + 0.5\, C_2\boldsymbol{g}(t-1),$$

and $\boldsymbol{S}^*$ satisfies the condition $\boldsymbol{S}^{*T}[1, \boldsymbol{g}^{*T}(t)]^T = 0$. It remains to show that $\boldsymbol{S}^* \succeq 0$ and $\lambda_2(\boldsymbol{S}^*) > 0$ with probability $1-o(1)$. This means

$$\mathbb{P}\Big(\inf_{\boldsymbol{V}\perp[1,\boldsymbol{g}^{*T}(t)]^T, \|\boldsymbol{V}\|=1} \boldsymbol{V}^T\boldsymbol{S}^*\boldsymbol{V} > 0\Big) \geq 1 - o(1), \quad (15)$$

where $\boldsymbol{V}$ is of dimension $n+1$. Let $\boldsymbol{V} \triangleq \left[v, \boldsymbol{U}^T\right]^T$, where v is a scalar and $\boldsymbol{U} \triangleq [u_1, u_2, \cdots, u_n]^T$. Therefore,

$$\begin{aligned}&\boldsymbol{V}^T\boldsymbol{S}^*\boldsymbol{V}\\ &= v^2\boldsymbol{S}_1^* - C_2 v\boldsymbol{U}^T\boldsymbol{g}(t-1) + \boldsymbol{U}^T\big(\boldsymbol{D}^* - C_1\boldsymbol{A}(t)\big)\boldsymbol{U}\\ &\geq (1-v^2)\Big[\min_{i\in[n]} d_i^* - C_1\big(\|\boldsymbol{A}(t) - \mathbb{E}[\boldsymbol{A}(t)]\| - p(1-2\xi)\big)\Big]\\ &+ v^2\Big[\frac{C_2}{2}\big(\boldsymbol{g}^T(t-1)\boldsymbol{g}^*(t) - 2\frac{\sqrt{n(1-v^2)}}{|v|}\big) - C_1 p(1-2\xi)\Big],\end{aligned}$$

where the inequality holds since

$$\begin{gathered}\boldsymbol{U}^T\boldsymbol{D}^*\boldsymbol{U} \geq (1-v^2)\min_{i\in[n]} d_i^*,\\ \boldsymbol{U}^T(\boldsymbol{A}(t) - \mathbb{E}[\boldsymbol{A}(t)])\boldsymbol{U} \leq (1-v^2)\|\boldsymbol{A}(t) - \mathbb{E}[\boldsymbol{A}(t)]\|,\\ v\boldsymbol{U}^T\boldsymbol{g}(t-1) \leq |v|\sqrt{n(1-v^2)}.\end{gathered}$$

Using the Chernoff bound, it can be proven that with probability $1 - o(1)$,

$$\sum_{i=1}^{n} g_i^*(t) g_i(t-1) \geq \sqrt{n}\ln n.$$

Thus, as $n \to \infty$,

$$v^2\Big[\frac{C_2}{2}\left(\sqrt{n}\ln n - 2\frac{\sqrt{n(1-v^2)}}{|v|}\right) - C_1 p(1-2\xi)\Big] \geq 0.$$

Applying [7, Theorem 9], for any c there exists some $c^{'} > 0$ such that for any $n \geq 1$,

$$\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\| \leq c^{'}\sqrt{\ln n}$$

with probability at least $1 - n^{-c}$, and hence

$$\boldsymbol{V}^T\boldsymbol{S}^*\boldsymbol{V} \geq (1-v^2)\Big[\min_{i\in[n]} d_i^* - C_1\big(c^{'}\sqrt{\ln n} - p(1-2\xi)\big)\Big]. \quad (16)$$

It follows from (14) that

$$\begin{aligned}\mathbb{P}(d_i^* \leq \delta) = \eta\mathbb{P}\Big(\sum_{j=1}^{n} A_{ij}(t) g_i^*(t) g_j^*(t) \leq \frac{\delta - 0.5\, C_2}{C_1}\Big)\\ + (1-\eta)\mathbb{P}\Big(\sum_{j=1}^{n} A_{ij}(t) g_i^*(t) g_j^*(t) \leq \frac{\delta + 0.5\, C_2}{C_1}\Big).\end{aligned}$$

where η is the probability of node label persistence in the Markov model. It can be shown that $\sum_{j=1}^{n} A_{ij}(t) g_i^*(t) g_j^*(t)$ has a distribution similar to $\sum_{i=1}^{n-1} S_i$ such that $\{S_1, S_2, ..., S_{n-1}\}$ is a sequence of i.i.d random variables with distribution $p_1\delta_{+1} + p_2\delta_{-1} + (1-p_1-p_2)\delta_0$, where $p_1 = \rho_1\frac{\ln n}{n}$ and $p_2 = \rho_2\frac{\ln n}{n}$ for positive ρ_1 and ρ_2.

Recall $\alpha \triangleq \lim_{n\to\infty}\frac{C_2}{2\ln n}$, $\beta \triangleq \sqrt{\alpha^2 + 4a^2C_1^2\xi(1-\xi)}$, and

$$\mu(a,\alpha) \triangleq a - \frac{\beta}{C_1} + \frac{\alpha}{2C_1}\ln\big(\frac{(1-\xi)(\alpha+\beta)}{\xi(\beta-\alpha)}\big).$$

First, we bound $\min_{i\in[n]} d_i^*$ under the condition $0 \leq \alpha < aC_1(1-2\xi)$. Using the union bound and [11, Lemma 9]

$$\mathbb{P}\Big(\min_{i\in[n]} d_i^* \geq \frac{\ln n}{\ln\ln n}\Big) \geq 1 - n^{1-\mu(a,\alpha)+o(1)}.$$

When $\mu(a,\alpha) > 1$, it follows $\min_{i\in[n]} d_i^* \geq \frac{\ln n}{\ln\ln n}$ with probability $1 - o(1)$. Thus, as long as $\mu(a,\alpha) > 1$, we can substitute $\min d_i^*$ in (16) with $\frac{\ln n}{\ln\ln n}$ and obtain

$$\begin{aligned}\boldsymbol{V}^T\boldsymbol{S}^*\boldsymbol{V} &\geq (1-v^2)\Big[\frac{\ln n}{\ln\ln n} - C_1\big(c^{'}\sqrt{\ln n} - p(1-2\xi)\big)\Big]\\ &> 0,\end{aligned}$$

which holds with probability $1 - o(1)$ as $n \to \infty$ and results in the first part of Theorem 1. For the second part, one should bound $\min_{i\in[n]} d_i^*$ under the condition that $\alpha > aC_1(1-2\xi)$. It follows from [11, Lemma 9] that

$$\mathbb{P}(d_i^* \leq \delta) \leq n^{-\mu(a,\alpha)+o(1)} + n^{-\alpha+o(1)},$$

where $\eta = 1 - n^{-\alpha+o(1)}$. Using the union bound,

$$\mathbb{P}\Big(\min_{i\in[n]} d_i^* \geq \delta\Big) \geq 1 - \Big(n^{1-\mu(a,\alpha)+o(1)} + n^{1-\alpha+o(1)}\Big).$$

Lemma 2. *If $\alpha > 1$, then $\mu(a,\alpha) > 1$.*

Proof. Define $\kappa(a,\alpha) \triangleq \mu(a,\alpha) - \alpha$. It can be proven that $\kappa(a,\alpha)$ is a convex function in α. Then

$$\mu(a,\alpha) - \alpha \geq a - \frac{\beta^*}{C_1}. \quad (17)$$

At the optimal α^*, $\beta^* = \alpha^* + 2\xi a C_1$ and

$$\mu(a,\alpha) - \alpha \geq a - 2\xi a - \frac{\alpha^*}{C_1}, \tag{18}$$

and $\mu^* = \frac{\alpha^*}{1-2\xi}$. This implies that $\alpha^* = (1-2\xi)aC_1$. Substituting in (18) leads to $\mu(a,\alpha) - \alpha \geq 0$, which implies that $\mu(a,\alpha) > 1$ when $\alpha > 1$. □

When $\alpha > 1$, Lemma 2 results in $\min_{i\in[n]} d_i^* \geq \frac{\ln n}{\ln\ln n}$ with probability $1-o(1)$. Substituting in (16), if $\alpha > 1$, with probability $1-o(1)$ one obtains

$$\begin{aligned} \boldsymbol{V}^T\boldsymbol{S}^*\boldsymbol{V} \geq\ & (1-v^2)\Big[\frac{\ln n}{\ln\ln n} - C_1\big(c'\sqrt{\ln n} - p(1-2\xi)\big)\Big] \\ >\ & 0, \end{aligned}$$

which proves the second part of Theorem 1.

B. Proof of Theorem 2

We must show that the maximum likelihood estimator fails. To do so, first let $T \triangleq \{\min_{i\in[n]}\ m_i^* \leq -C_1\}$, where

$$m_i^* = C_1 \sum_{j=1}^{n} A_{ij}(t)\, g_j^*(t)\, g_i^*(t) + 0.5\, C_2 g_i(t-1) g_i^*(t).$$

Then, $\mathbb{P}(\text{maximum likelihood Fails}) \geq \mathbb{P}(T)$.

Let K denotes the set of first $\lfloor \frac{n}{\ln^2 n} \rfloor$ nodes, $e(i,K)$ denotes the number of edges between node i and nodes in the set $K \subset [n]$, and

$$h_i^* = \sum_{j\in K^c} A_{ij}(t) g_i^*(t) g_j^*(t).$$

It can be shown that

$$\begin{aligned} &\min_{i\in[n]}\ m_i^* \leq \min_{i\in K} m_i^* \\ &\leq \min_{i\in K}\ \Big(C_1 h_i^* + 0.5\, C_2 g_i^*(t) g_i(t-1)\Big) + \max_{i\in K}\ e(i,K). \end{aligned}$$

Define

$$\begin{aligned} E_1 &\triangleq \{\max_{i\in K}\ e(i,K) \leq \delta - C_1\}, \\ E_2 &\triangleq \big\{\min_{i\in K}\ \big(C_1 h_i^* + 0.5\, C_2 g_i^*(t) g_i(t-1)\big) \leq -\delta\big\}. \end{aligned}$$

Since $T \supset E_1 \cap E_2$, it suffices to show $\mathbb{P}(E_1) \to 1$ and $\mathbb{P}(E_2) \to 1$ to prove that the maximum likelihood estimator fails.

Since $e(i,K) \sim \text{Binom}(|K|, a\frac{\ln n}{n})$, from [11, Lemma 9],

$$\begin{aligned} &\mathbb{P}(e(i,K) \geq \delta - C_1) \\ &\leq \left(\frac{\ln^2 n}{a e \ln\ln n} - \frac{C_1 \ln n}{a e}\right)^{C_1 - \frac{\ln n}{\ln\ln n}} e^{-\frac{a}{\ln n}} \leq n^{-2+o(1)}. \end{aligned}$$

Using the union bound, $\mathbb{P}(E_1) \geq 1 - n^{-1+o(1)}$. Let

$$\begin{aligned} E &\triangleq \Big\{C_1 h_i^* + 0.5\, C_2 g_i^*(t) g_i(t-1) \leq -\delta\Big\}, \\ E_+ &\triangleq \Big\{h_i^* \leq \frac{-\delta - 0.5\, C_2}{C_1}\Big\} \\ E_- &\triangleq \Big\{h_i^* \leq \frac{-\delta + 0.5\, C_2}{C_1}\Big\}. \end{aligned}$$

Then

$$\begin{aligned} \mathbb{P}(E_2) &= 1 - \prod_{i\in K}[1-\mathbb{P}(E)] \overset{(a)}{=} 1 - [1-\mathbb{P}(E)]^{|K|} \\ &= 1 - \big[1 - (1-\eta)\mathbb{P}(E_-) - \eta\mathbb{P}(E_+)\big]^{|K|}, \end{aligned} \tag{19}$$

where (a) holds because $\{C_1 h_i^* + 0.5\, C_2 g_i^*(t) g_i(t-1)\}_{i\in K}$ are mutually independent.

First, we bound $\mathbb{P}(E_2)$ under the condition $0 \leq \alpha < aC_1(1-2\xi)$. Applying [11, Lemma 9],

$$\begin{aligned} \mathbb{P}(E_+) &\geq n^{-\mu(a,\alpha)+o(1)}, \\ \mathbb{P}(E_-) &\geq n^{-\mu(a,\alpha)+\alpha+o(1)} \end{aligned}$$

It follows from (19) that

$$\mathbb{P}(E_2) \geq 1 - \exp\Big(-n^{1-\mu(a,\alpha)+o(1)}\Big), \tag{20}$$

where the inequality holds because $\eta = 1 - n^{-\alpha+o(1)}$ and $1 + x \leq e^x$. Therefore, if $\mu(a,\alpha) < 1$, then $\mathbb{P}(E_2) \to 1$ and the first part of Theorem 2 follows.

We now find the bounds of $\mathbb{P}(E_2)$ under $\alpha \geq aC_1(1-2\xi)$. Reorganizing (19),

$$\mathbb{P}(E_2) = 1 - \big[\eta\mathbb{P}(E_+^c) + (1-\eta)\mathbb{P}(E_-^c)\big]^{|K|}, \tag{21}$$

where

$$\begin{aligned} \mathbb{P}(E_+^c) &= \mathbb{P}\bigg(h_i^* \geq \frac{-\delta - 0.5\, C_2}{C_1}\bigg), \\ \mathbb{P}(E_-^c) &= \mathbb{P}\bigg(h_i^* \geq \frac{-\delta + 0.5\, C_2}{C_1}\bigg). \end{aligned}$$

Also, h_i^* is equal in distribution to $\sum_{i=1}^{|K^c|-1} S_i$ according to [11, Lemma 9], where $p_1 \triangleq p(1-\xi)$ and $p_2 \triangleq p\xi$. Then, $\mathbb{P}(E_+^c) \leq 1$ and

$$\mathbb{P}(E_-^c) \leq n^{-\mu(a,\alpha)+\alpha+o(1)}.$$

It follows from (21) that

$$\begin{aligned} \mathbb{P}(E_2) &\geq 1 - \Big[\eta + (1-\eta) n^{-\mu(a,\alpha)+\alpha+o(1)}\Big]^{|K|} \\ &\overset{(a)}{=} 1 - \Big[1 - n^{-\alpha+o(1)} + n^{-\mu(a,\alpha)+o(1)}\Big]^{|K|} \\ &\overset{(b)}{\geq} 1 - exp\Big(-n^{1-\alpha+o(1)}\Big(1 - n^{-\mu(a,\alpha)+\alpha+o(1)}\Big)\Big) \end{aligned}$$

where (a) holds because $\eta = 1 - n^{-\alpha+o(1)}$ and (b) is due to $1 + x < exp(x)$. Therefore, since $\alpha \leq \mu(a,\alpha)$, if $\alpha < 1$, then $\mathbb{P}(E_2) \to 1$ and the second part of Theorem 2 follows.

VI. Conclusions

This paper addresses community detection in the dynamic binary CBM, considering exact recovery metric. Under this model, it is assumed that the node labels of the current snapshot are dependent on the previous τ sequential snapshots. The maximum likelihood estimator is extracted for this problem. Relaxing the maximum likelihood estimator, a semidefinite programming algorithm is formulated to exactly recover the current node labels. A sharp asymptotic threshold is proven for the exact recovery under certain regimes for the investigated dynamic binary CBM.

References

[1] P. Wagenseller, F. Wang, and W. Wu, "Size matters: A comparative analysis of community detection algorithms," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 4, pp. 951–960, 2018.

[2] M. Esmaeili, H. Saad, and A. Nosratinia, "Exact recovery by semidefinite programming in the binary stochastic block model with partially revealed side information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3477–3481.

[3] Q. Zhang, V. Y. F. Tan, and C. Suh, "Community detection and matrix completion with social and item similarity graphs," *IEEE Transactions on Signal Processing*, vol. 69, pp. 917–931, 2021.

[4] T. Ji, C. Luo, Y. Guo, Q. Wang, L. Yu, and P. Li, "Community detection in online social networks: A differentially private and parsimonious approach," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 1, pp. 151–163, 2020.

[5] M. Li, X. Meng, R. Zheng, F.-X. Wu, Y. Li, Y. Pan, and J. Wang, "Identification of protein complexes by using a spatial and temporal active protein interaction network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 3, pp. 817–827, 2020.

[6] L. Huang, Y. Yang, H. Gao, X. Zhao, and Z. Du, "Comparing community detection algorithms in transport networks via points of interest," *IEEE Access*, vol. 6, pp. 29 729–29 738, 2018.

[7] B. Hajek, Y. Wu, and J. Xu, "Achieving exact cluster recovery threshold via semidefinite programming," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2788–2797, 2016.

[8] E. Abbe, "Community detection and stochastic block models: Recent developments," *Journal of Machine Learning Research*, vol. 18, no. 177, pp. 1–86, 2018.

[9] L. Dall'Amico, R. Couillet, and N. Tremblay, "Optimal laplacian regularization for sparse spectral community detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3237–3241.

[10] H. Saad and A. Nosratinia, "Community detection with side information: Exact recovery under the stochastic block model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 944–958, 2018.

[11] M. Esmaeili, H. M. Saad, and A. Nosratinia, "Semidefinite programming for community detection with side information," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 2, pp. 1957–1973, 2021.

[12] M. Esmaeili, H. Saad, and A. Nosratinia, "Community detection with side information via semidefinite programming," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 420–424.

[13] M. Esmaeili and A. Nosratinia, "Community detection: Exact recovery in weighted graphs," 2021. [Online]. Available: https://arxiv.org/abs/2102.04439

[14] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer, "Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery," *IEEE Transactions on Network Science and Engineering*, vol. 1, no. 1, pp. 10–22, 2014.

[15] S. Dhara, J. Gaudio, E. Mossel, and C. Sandon, "Spectral recovery of binary censored block models," 2021. [Online]. Available: https://arxiv.org/abs/2107.06338

[16] M. Esmaeili and A. Nosratinia, "Community detection with secondary latent variables," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 1355–1360.

[17] K. S. Xu and A. O. Hero, "Dynamic stochastic blockmodels for time-evolving social networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 552–562, 2014.

[18] A. Ghasemian, P. Zhang, A. Clauset, C. Moore, and L. Peel, "Detectability thresholds and optimal algorithms for community structure in dynamic networks," *Phys. Rev. X*, vol. 6, p. 031005, Jul 2016.

[19] P. Barucca, F. Lillo, P. Mazzarisi, and D. Tantari, "Disentangling group and link persistence in dynamic stochastic block models," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2018, no. 12, p. 123407, dec 2018.

[20] M. Pensky and T. Zhang, "Spectral clustering in the dynamic stochastic block model," *Electronic Journal of Statistics*, vol. 13, no. 1, pp. 678 – 709, 2019.

[21] L. Dall'Amico, R. Couillet, and N. Tremblay, "Community detection in sparse time-evolving graphs with a dynamical bethe-hessian," 2020. [Online]. Available: https://arxiv.org/abs/2006.04510

[22] E. Mossel and J. Xu, "Local algorithms for block models with side information," ser. ITCS '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 71–80.