The Annals of Statistics
2019, Vol. 47, No. 4, 2378–2403
https://doi.org/10.1214/18-AOS1751
© Institute of Mathematical Statistics, 2019

DYNAMIC NETWORK MODELS AND GRAPHON ESTIMATION

BY MARIANNA PENSKY[1]

University of Central Florida

In the present paper, we consider a dynamic stochastic network model. The objective is estimation of the tensor of connection probabilities $\mathbf{\Lambda}$ when it is generated by a Dynamic Stochastic Block Model (DSBM) or a dynamic graphon. In particular, in the context of the DSBM, we derive a penalized least squares estimator $\widehat{\mathbf{\Lambda}}$ of $\mathbf{\Lambda}$ and show that $\widehat{\mathbf{\Lambda}}$ satisfies an oracle inequality and also attains minimax lower bounds for the risk. We extend those results to estimation of $\mathbf{\Lambda}$ when it is generated by a dynamic graphon function. The estimators constructed in the paper are adaptive to the unknown number of blocks in the context of the DSBM or to the smoothness of the graphon function. The technique relies on the vectorization of the model and leads to much simpler mathematical arguments than the ones used previously in the stationary set up. In addition, all results in the paper are nonasymptotic and allow a variety of extensions.

1. Introduction. Networks arise in many areas of research such as sociology, biology, genetics, ecology, information technology to list a few. An overview of statistical modeling of random graphs can be found in, for example, Kolaczyk (2009) and Goldenberg et al. (2010). While static network models are relatively well understood, the literature on the dynamic network models is fairly recent.

In this paper, we consider a dynamic network defined as an undirected graph with n nodes with connection probabilities changing in time. Assume that we observe the values of a tensor $\mathbf{B}_{i,j,l} \in \{0,1\}$ at times t_l where $0 < t_1 < \cdots < t_L = T$. For simplicity, we assume that time instants are equispaced and the time interval is scaled to one, that is, $t_l = l/L$. Here, $\mathbf{B}_{i,j,l} = 1$ if a connection between nodes i and j is observed at time t_l and $\mathbf{B}_{i,j,l} = 0$ otherwise. We set $\mathbf{B}_{i,i,l} = 0$ and $\mathbf{B}_{i,j,l} = \mathbf{B}_{j,i,l}$ for any $i, j = 1, \dots, n$ and $l = 1, \dots, L$, and assume that $\mathbf{B}_{i,j,l}$ are independent Bernoulli random variables with $\mathbf{\Lambda}_{i,j,l} = \mathbb{P}(\mathbf{B}_{i,j,l} = 1)$ and $\mathbf{\Lambda}_{i,i,l} = 0$. Below, we study two types of objects: a Dynamic Stochastic Block Model (DSBM) and a dynamic graphon.

The DSBM can be viewed as a natural extension of the Stochastic Block Model (SBM) which, according to Olhede and Wolfe (2014), provides a universal tool for description of time-independent stochastic network data. In a DSBM, all n

Received April 2017; revised March 2018.

[1]Supported in part by NSF Grants DMS-14-07475 and DMS-17-12977.

MSC2010 subject classifications. Primary 60G05; secondary 05C80, 62F35.

Key words and phrases. Dynamic network, graphon, stochastic block model, nonparametric regression, minimax rate.

nodes are grouped into m classes $\Omega_1, \dots, \Omega_m$, and probability of a connection $\mathbf{\Lambda}_{i,j,l}$ is entirely determined by the groups to which the nodes i and j belong at the moment t_l. In particular, if $i \in \Omega_k$ and $j \in \Omega_{k'}$, then $\mathbf{\Lambda}_{i,j,l} = \mathbf{G}_{k,k',l}$. Here, $\mathbf{G}$ is the *connectivity tensor* at time t_l with $\mathbf{G}_{k,k',l} = \mathbf{G}_{k',k,l}$. Denote by $n_k^{(l)}$ the number of nodes in class k at the moment t_l, $k = 1, \dots, m$, $l = 1, \dots, L$.

A dynamic graphon can be defined as follows. Let $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)$ be a random vector sampled from a distribution $\mathbb{P}_\zeta$ supported on $[0,1]^n$. Although the most common choice for $\mathbb{P}_\zeta$ is the i.i.d. uniform distribution for each ζ_i, we do not make this assumption in the present paper. We further assume that there exists a function $f : [0,1]^3 \to [0,1]$ such that for any t one has $f(x,y,t) = f(y,x,t)$ and

$$\mathbf{\Lambda}_{i,j,l} = f(\zeta_i, \zeta_j, t_l), \qquad i, j = 1, \dots, n, l = 1, \dots, L. \tag{1}$$

Then function f summarizes behavior of the network and can be called *dynamic graphon*, similarly to the graphon in the situation of a stationary network. This formulation allows to study a different set of stochastic network models than the DSBM.

It is known that graphons play an important role in the theory of graph limits described in Lovász and Szegedy (2006) and Lovász (2012). The definition of the dynamic graphon above fully agrees with their theory. Indeed, for every $l = 1, \dots, L$, the limit of $\mathbf{\Lambda}_{*,*,l}$ as $n \to \infty$ is $f(\cdot, \cdot, t_l)$. We shall further elaborate on the notion of the dynamic graphon in Section 7.

In the last few years, dynamic network models attracted a great deal of attention [see, e.g., Durante, Dunson and Vogelstein (2017), Durante and Dunson (2016), Han, Xu and Airoldi (2015), Kolar et al. (2010), Anagnostopoulos et al. (2016), Matias and Miele (2017), Minhas, Hoff and Warda (2015), Xing, Fu and Song (2010), Xu (2015), Xu and Hero III (2014) and Yang et al. (2011) among others]. The majority of those papers describe changes in the connection probabilities and group memberships via various kinds of Bayesian or Markov random field models and carry out the inference using the EM or iterative optimization algorithms. While procedures described in those papers show good computational properties, they come without guarantees for the estimation precision. The only paper known to us that is concerned with estimation precision in the dynamic setting is by Han, Xu and Airoldi (2015) where the authors study consistency of their procedures when $n \to \infty$ or $L \to \infty$.

On the other hand, recently, several authors carried out minimax studies in the context of stationary network models. In particular, Gao, Lu and Zhou (2015) developed upper and minimax lower bounds for the risk of estimation of the matrix of connection probabilities. In a subsequent paper, Gao et al. (2016) generalized the results to a somewhat more general problem of estimation of matrices with biclustering structures. In addition, Klopp, Tsybakov and Verzelen (2017) extended these results to the case when the network is sparse in a sense that probability of connection is uniformly small and tends to zero as $n \to \infty$. Also, Zhang and

Zhou (2016) investigated minimax rates of community detection in the two-class stochastic block model.

The present paper has several objectives. First, we describe the non-parametric DSBM model that allows for smooth evolution of the tensor $\mathbf{G}$ of connection probabilities as well as changes in group memberships in time. Second, we introduce vectorization of the model that enables us to take advantage of well-studied methodologies in nonparametric regression estimation. Using these techniques, we derive penalized least squares estimators $\widehat{\mathbf{\Lambda}}$ of $\mathbf{\Lambda}$ and show that they satisfy oracle inequalities. These inequalities do not require any assumptions on the mechanism that drives evolution of the group memberships of the nodes in time and can be applied under very mild conditions. Furthermore, we consider a particular situation where only at most n_0 nodes can change their memberships between two consecutive time points. Under the latter assumption, we derive minimax lower bounds for the risk of an estimator of $\mathbf{\Lambda}$ and confirm that the estimators constructed in the paper attain those lower bounds. Moreover, we extend those results to estimation of the tensor $\mathbf{\Lambda}$ when it is generated by a graphon function. We show that, for the graphon, the estimators are minimax optimal within a logarithmic factor of L. Estimators, constructed in the paper, do not require knowledge of the number of classes m in the context of the DSBM, or a degree of smoothness of the graphon function f if $\mathbf{\Lambda}$ is generated by a dynamic graphon.

Note that unlike in Klopp, Tsybakov and Verzelen (2017) we do not consider a network that is sparse in a sense that probabilities of connections between classes are uniformly small. However, since our technique is based on model selection, it allows to study a network where some groups do not communicate with each other and obtain more accurate results. Moreover, as we show in Section 6, by adjusting the penalty, one can provide adaptation to uniform sparsity assumption if the number of nodes in each class is large enough.

The present paper makes several key contributions. First, to the best of our knowledge, the time-dependent networks are usually handled via generative models that assume some probabilistic mechanism which governs the evolution of the network in time. The present paper offers the first fully nonparametric model for the time-dependent networks which does not make any of such assumptions. It treats connection probabilities for each group as functional data, allows group membership switching and enables one to exploit stability in the group memberships over time. Second, the paper provides the first minimax study of estimation of the tensor of connection probabilities in a dynamic setting. The estimators constructed in the paper are adaptive to the number of blocks in the context of the DSBM and to the smoothness of the graphon function in the case of a dynamic graphon. Moreover, the approach of the paper is nonasymptotic, so it can be used irrespective of how large the number of nodes n, the number of groups m and the number of time instants L are and what the relationship between these parameters is. Third, in order to handle the tensor-variate functional data, we use vectorization of the model. This technique allows to reduce the problem of estimation of

an unknown tensor of connection probabilities to a solution of a functional linear regression problem with sub-Gaussian errors. The technique is very potent and is used in a novel way. In particular, it leads to much more simple mathematics than in Gao, Lu and Zhou (2015) and Klopp, Tsybakov and Verzelen (2017). In the case of a time-independent SBM, it immediately reduces the SBM to a linear regression setting. In addition, by using the properties of the Kronecker product, we are able to reduce the smoothness assumption on the connection probabilities to sparsity assumption on their coefficients in one of the common orthogonal transforms (e.g., Fourier or wavelet). Fourth, we use the novel structure of the penalty a part of which is proportional to the logarithm of the cardinality of the set of all possible clustering matrices over L time instants. The latter allows to accommodate various group membership switching scenarios and is based on the Packing lemma (Lemma 4) which can be viewed as a version of the Varshamov–Gilbert lemma for clustering matrices. In particular, while all papers that studied the SBM dealt with the situation where no restrictions are placed on the set of clustering matrices, our approach allows to impose those restrictions. Finally, the methodologies of the paper admit various generalizations. For example, they can be adapted to a situation where the number of nodes in the network depends on time, or the connection probabilities have jump discontinuities, or when some of the groups have no connection with each other. Section 6 shows that the technique can be adapted to an additional uniform sparsity considered in Klopp, Tsybakov and Verzelen (2017) if the number of nodes in each class is large enough.

The rest of the paper is organized as follows. In Section 2, we introduce notation and describe the vectorization of the model. In Section 3, we construct the penalized least squares estimators $\widehat{\mathbf{\Lambda}}$ of the tensor $\mathbf{\Lambda}$. In Section 4, we derive the oracle inequalities for their risks. In Section 5, we obtain the minimax lower bounds for the risk that confirm that the estimators $\widehat{\mathbf{\Lambda}}$ are minimax optimal. Section 6 shows how our technique provides adaptation to uniform sparsity assumption studied in Klopp, Tsybakov and Verzelen (2017). Section 7 develops the nearly minimax optimal (within a logarithmic factor of L) estimators of $\mathbf{\Lambda}$ when the network is generated by a graphon. Finally, Section 8, provides a discussion of various generalizations of the techniques proposed in the paper. The proofs of all statements are placed into the Supplementary Material [Pensky (2019)].

2. Notation, discussion of the model and data structures.

2.1. *Notation.* For any two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ means that there exists a constant $C > 0$ independent of n such that $C^{-1}a_n \le b_n \le Ca_n$ for any n. For any set Ω, denote cardinality of Ω by $|\Omega|$. For any x, $[x]$ is the largest integer no larger than x.

For any vector $\mathbf{t} \in \mathbb{R}^p$, denote its ℓ_2, ℓ_1, ℓ_0 and ℓ_∞ norms by, respectively, $\|\mathbf{t}\|$, $\|\mathbf{t}\|_1$, $\|\mathbf{t}\|_0$ and $\|\mathbf{t}\|_\infty$. Denote by $\|\mathbf{t}_1 - \mathbf{t}_2\|_H$ the Hamming distance between vectors $\mathbf{t}_1$ and $\mathbf{t}_2$. Denote by $\mathbf{1}$ and $\mathbf{0}$ the vectors that have, respectively, only unit or zero

elements. Denote by $\mathbf{e}_j$ the vector with 1 in the jth position and all other elements equal to zero.

For a matrix $\mathbf{A}$, its ith row and jth columns are denoted, respectively, by $\mathbf{A}_{i,*}$ and $\mathbf{A}_{*,j}$. Similarly, for a tensor $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we denote its lth $(n_1 \times n_2)$-dimensional submatrix by $\mathbf{A}_{*,*,l}$. Let $\operatorname{vec}(\mathbf{A})$ be the vector obtained from matrix $\mathbf{A}$ by sequentially stacking its columns. Denote by $\mathbf{A} \otimes \mathbf{B}$ the Kronecker product of matrices $\mathbf{A}$ and $\mathbf{B}$. Also, $\mathbf{I}_k$ is the identity matrix of size k. For any subset J of indices, any vector $\mathbf{t}$ and any matrix $\mathbf{A}$, denote the restriction of $\mathbf{t}$ to indices in J by $\mathbf{t}_J$ and the restriction of $\mathbf{A}$ to columns $\mathbf{A}_{*,j}$ with $j \in J$ by $\mathbf{A}_J$. Also, denote by $\mathbf{t}_{(J)}$ the modification of vector $\mathbf{t}$ where all elements $\mathbf{t}_j$ with $j \notin J$ are set to zero.

For any matrix $\mathbf{A}$, denote its spectral and Frobenius norms by, respectively, $\|\mathbf{A}\|_{op}$ and $\|\mathbf{A}\|$. Denote $\|\mathbf{A}\|_H \equiv \|\operatorname{vec}(\mathbf{A})\|_H$, $\|\mathbf{A}\|_\infty = \|\operatorname{vec}(\mathbf{A})\|_\infty$ and $\|\mathbf{A}\|_0 \equiv \|\operatorname{vec}(\mathbf{A})\|_0$. For any tensor $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, denote $\|\mathbf{A}\|^2 = \sum_{l=1}^{n_3} \|\mathbf{A}_{*,*,l}\|^2$.

Denote by $\mathcal{M}(m,n)$ a collection of *membership* (or *clustering*) matrices $\mathbf{Z} \in \{0,1\}^{n \times m}$, that is, matrices such that $\mathbf{Z}$ has exactly one 1 per row and $\mathbf{Z}_{ik} = 1$ iff a node i belongs to the class Ω_k and is zero otherwise. Denote by $\mathcal{C}(m,n,L)$ a set of clustering matrices such that

$$\mathcal{C}(m,n,L) \subseteq \prod_{l=1}^{L} \mathcal{M}(m,n). \tag{2}$$

2.2. *Discussion of the model.* Note that the values of $\mathbf{B}_{i,j,l}$ are independent given the values of $\mathbf{\Lambda}_{i,j,l}$, that is, $\mathbf{B}_{i,j,l}$ are independent in the sense that their deviations from $\mathbf{\Lambda}_{i,j,l}$ are independent from each other. Therefore, the values of $\mathbf{B}_{i,j,l}$ are linked to each other in the same way as observations of a continuous function with independent Gaussian errors are related to each other. Moreover, in majority of papers treating dynamic block models [see, e.g., Durante, Dunson and Vogelstein (2017), Han, Xu and Airoldi (2015), Matias and Miele (2017), Yang et al. (2011) among others], similar to the present paper, the authors assume that observations $\mathbf{B}_{i,j,l}$ are independent given $\mathbf{\Lambda}_{i,j,l}$. Note that this is not an artificial construct: Durante, Dunson and Vogelstein (2017), for example, use the model for studying international relationships between countries over time.

The only difference between the present paper and the papers cited above is that we assume that the underlying connection probabilities $\mathbf{G}_{*,*,l}$ are functionally linked (e.g., smooth) rather than being probabilistically related. Indeed, many papers that treat dynamic block models assume some Bayesian generative mechanism on the values of connection probabilities as well as on evolution of clustering matrices. In particular, they impose some prior distributions that relate $\mathbf{G}_{*,*,l+1}$ to $\mathbf{G}_{*,*,l}$ and $\tilde{\mathbf{Z}}^{(l+1)}$ to $\tilde{\mathbf{Z}}^{(l)}$, the matrices of underlying probabilities and the clustering matrices for consecutive time points. Since the proposed generative mechanism may be invalid, we avoid making assumptions about the probabilistic structures that generate connection probabilities and group memberships, and treat the network as a given object. However, our model enforces, in a sense, a more close but

yet flexible relation between the values of $\mathbf{B}_{i,j,l}$ since $\mathbf{G}_{*,*,l}$ are functionally (and not stochastically) related. Moreover, our theory allows to place any restrictions on the set of clustering matrices.

To illustrate this point, consider just one pair of nodes (i, j) and assume that these nodes do not switch their memberships between times t_l and t_{l+1} and also that $\mathbf{G}_{i,j,l}$ is continuous at t_l. It is easy to see that if $\mathbf{G}_{i,j,l}$ is close to zero (or one), then $\mathbf{G}_{i,j,l+1}$ is also close to zero (or one), and hence, $\mathbf{B}_{i,j,l}$ and $\mathbf{B}_{i,j,l+1}$ are likely to be equal to zero (or one) simultaneously. This relationship takes place in general.

To simplify the narrative, just for this paragraph, denote $b_l = \mathbf{B}_{i,j,l}$, $b_{l+1} = \mathbf{B}_{i,j,l+1}$, $g_l = \mathbf{G}_{i,j,l}$ and $g_{l+1} = \mathbf{G}_{i,j,l+1}$. In order we are able to assert conditional probabilities $\mathbb{P}(\mathbf{B}_{i,j,l+1} = 1|\mathbf{B}_{i,j,l} = 1) \equiv \mathbb{P}(b_{l+1} = 1|b_l = 1)$ and $\mathbb{P}(\mathbf{B}_{i,j,l+1} = 0|\mathbf{B}_{i,j,l} = 0) \equiv \mathbb{P}(b_{l+1} = 0|b_l = 0)$, consider the situation where g_l and g_{l+1} are random variables with the joint pdf $p(g_l, g_{l+1})$ such that, given g_l, on the average g_{l+1} is equal to g_l: $\mathbb{E}(g_{l+1} \mid g_l) = g_l$. Assume, as it is done in the present paper, that given g_l, values of b_l are independent Bernoulli variables, so that

$$p(b_l, b_{l+1} \mid g_l, g_{l+1}) = g_l^{b_l}(1 - g_l)^{1-b_l} g_{l+1}^{b_{l+1}}(1 - g_{l+1})^{1-b_{l+1}}.$$

It is straightforward to calculate marginal probabilities $\mathbb{P}(b_l = 1) = \mathbb{E}(g_l)$, $\mathbb{P}(b_{l+1} = 1) = \mathbb{E}(g_{l+1}) = \mathbb{E}[\mathbb{E}(g_{l+1} \mid g_l)] = \mathbb{E}(g_l)$ and the joint probability $\mathbb{P}(b_l = 1, b_{l+1} = 1) = \mathbb{E}(g_{l+1}g_l) = \mathbb{E}[\mathbb{E}(g_{l+1}g_l \mid g_l)] = \mathbb{E}(g_l^2)$ which yields

$$\mathbb{P}(b_{l+1} = 1 \mid b_l = 1) - \mathbb{P}(b_{l+1} = 1) = \frac{\mathbb{E}(g_l^2)}{\mathbb{E}(g_l)} - \mathbb{E}(g_l) = \frac{\mathrm{Var}(g_l)}{\mathbb{E}(g_l)} > 0$$

unless $\mathrm{Var}(g_l) = 0$. The latter means that, even in the presence of the assumption of the conditional independence, the probability of interaction at the moment t_{l+1} is larger if there were an interaction at the moment t_l than it would be in the absence of this assumption. Similarly, repeating the calculation with g_l and g_{l+1} replaced by $1 - g_l$ and $1 - g_{l+1}$, obtain

$$\mathbb{P}(b_{l+1} = 0 \mid b_l = 0) - \mathbb{P}(b_{l+1} = 0) = \frac{\mathrm{Var}(g_l)}{\mathbb{E}(1 - g_l)} > 0.$$

In the absence of the probabilistic assumptions on g_l and g_{l+1}, we cannot evaluate those conditional probabilities but the relationship persists in this situation as well.

2.3. *Vectorization of the model*. Note that tensor $\mathbf{\Lambda}$ of connection probabilities has a lot of structure. On one hand, it is easy to check that

$$\mathbf{\Lambda}_{*,*,l} = \tilde{\mathbf{Z}}^{(l)}\mathbf{G}_{*,*,l}\big(\tilde{\mathbf{Z}}^{(l)}\big)^T, \qquad \mathbf{B}_{i,j,l} \sim \mathrm{Bernoulli}(\mathbf{\Lambda}_{i,j,l}), \tag{3}$$

where $\tilde{\mathbf{Z}}^{(l)} \in \mathcal{M}(m, n)$ is the clustering matrix at the moment t_l. On the other hand, for every k_1 and k_2, vectors $\mathbf{G}_{k_1,k_2,*} \in \mathbb{R}^L$ are comprised of values of some smooth functions and, therefore, have low complexity. Usually, efficient representations of

such vectors are achieved by applying some orthogonal transform $\mathbf{H}$ (e.g., Fourier or wavelet transform); however, we cannot apply this transform to the original data tensor for two reasons. First, the errors in the model are not Gaussian, so application of $\mathbf{H}$ will convert the data tensor with independent Bernoulli components into a data tensor with dependent entries that are not Bernoulli variables any more. In addition, application of this transform to the original data will not achieve our goals since, although vectors $\mathbf{G}_{k_1,k_2,*}$ represent smooth functions, vectors $\mathbf{\Lambda}_{i,j,*}$ do not, due to possible switches in the group memberships. In addition, for every l, matrix $\mathbf{\Lambda}_{*,*,l}$ in (3) forms the so called bi-clustering structure [see, e.g., Gao et al. (2016)] which makes recovery of $\mathbf{G}_{*,*,l}$ much harder than in the case of a usual regression model.

In order to handle all these intrinsic difficulties, we apply operation of vectorization to $\mathbf{\Lambda}_{*,*,l}$. Denote

$$\boldsymbol{\lambda}^{(l)} = \operatorname{vec}(\mathbf{\Lambda}_{*,*,l}), \qquad \mathbf{b}^{(l)} = \operatorname{vec}(\mathbf{B}_{*,*,l}), \qquad \mathbf{g}^{(l)} = \operatorname{vec}(\mathbf{G}_{*,*,l}). \tag{4}$$

Then Theorem 1.2.22(i) of Gupta and Nagar (2000) yields

$$\begin{aligned} &\boldsymbol{\lambda}^{(l)} = (\tilde{\mathbf{Z}}^{(l)} \otimes \tilde{\mathbf{Z}}^{(l)})\mathbf{g}^{(l)}, \\ &\mathbf{b}_i^{(l)} \sim \text{Bernoulli}(\boldsymbol{\lambda}_i^{(l)}), \qquad i = 1, \dots, n^2, l = 1, \dots, L. \end{aligned} \tag{5}$$

Note that $\mathbf{b}_i^{(l)}$ in (5) are independent for different values of l but not i due to the symmetry. In addition, the values of $\mathbf{b}_i^{(l)}$ and $\boldsymbol{\lambda}_i^{(l)}$ that are corresponding to diagonal elements of matrices $\mathbf{B}_{*,*,l}$ and $\mathbf{\Lambda}_{*,*,l}$, are equal to zero by construction. Since all those values are not useful for estimation, we remove redundant entries from vectors $\boldsymbol{\lambda}^{(l)}$ and $\mathbf{b}^{(l)}$ for every $l = 1, \dots, L$. Specifically, in (5), we remove the elements in $\boldsymbol{\lambda}^{(l)}$ and the rows in $(\tilde{\mathbf{Z}}^{(l)} \otimes \tilde{\mathbf{Z}}^{(l)})$ corresponding, respectively, to $\mathbf{\Lambda}_{i_1,i_2,l}$ and $(\tilde{\mathbf{Z}}_{i_1,*}^{(l)} \otimes \tilde{\mathbf{Z}}_{i_2,*}^{(l)})$ with $i_1 \geq i_2$. We denote the reductions of vectors $\boldsymbol{\lambda}^{(l)}$, $\mathbf{b}^{(l)}$ and matrices $(\tilde{\mathbf{Z}}^{(l)} \otimes \tilde{\mathbf{Z}}^{(l)})$ by, respectively, $\boldsymbol{\theta}^{(l)}$, $\mathbf{a}^{(l)}$ and $\tilde{\mathbf{C}}^{(l)}$ obtaining

$$\begin{aligned} &\boldsymbol{\theta}^{(l)} = \tilde{\mathbf{C}}^{(l)}\mathbf{g}^{(l)}, \\ &\mathbf{a}_i^{(l)} \sim \text{Bernoulli}(\boldsymbol{\theta}_i^{(l)}), \qquad i = 1, \dots, n(n-1)/2, l = 1, \dots, L. \end{aligned} \tag{6}$$

Note that unlike in the case of $\mathbf{b}^{(l)}$, elements $\mathbf{a}_i^{(l)}$ and $\mathbf{a}_{i'}^{(l')}$ are independent whenever $i \neq i'$ or $l \neq l'$. The interesting thing here is that matrices $\tilde{\mathbf{C}}^{(l)}$ are still clustering matrices, that is, $\tilde{\mathbf{C}}^{(l)} \in \mathcal{M}(n(n-1)/2, m^2)$. Indeed, $\tilde{\mathbf{C}}^{(l)}$ are binary matrices such that, for i corresponding to (i_1, i_2) with $i_1 < i_2$ and k corresponding to (k_1, k_2) in $(\tilde{\mathbf{Z}}_{i_1,k_1}^{(l)} \otimes \tilde{\mathbf{Z}}_{i_2,k_2}^{(l)})$ one has $\tilde{\mathbf{C}}_{i,k}^{(l)} = 1$ if and only if the nodes $i_1 \in \Omega_{k_1}$ and $i_2 \in \Omega_{k_2}$.

Observe that although we removed the redundant elements from vectors $\boldsymbol{\lambda}^{(l)}$ and $\mathbf{b}^{(l)}$, we have not done so for the vectors $\mathbf{g}^{(l)}$. Indeed, since matrices $\mathbf{G}_{*,*,l}$ are symmetric, the elements of vectors $\mathbf{g}^{(l)}$ corresponding to $\mathbf{G}_{k_1,k_2,l}$ and $\mathbf{G}_{k_2,k_1,l}$

with $k_1 \neq k_2$ are equal to each other. For the sake of eliminating such redundancy (and hence, the need of tracing the equal elements in the process of estimation), for indices k corresponding to pairs of classes (k_1, k_2) with $k_1 > k_2$, we remove entries $\mathbf{g}_k^{(l)}$ from vectors $\mathbf{g}^{(l)}$ and denote the resulting vectors by $\mathbf{q}^{(l)}$. In order an equivalent of the relation (6) still holds with vectors $\mathbf{q}^{(l)}$ instead of $\mathbf{g}^{(l)}$, we add together columns of matrices $\tilde{\mathbf{C}}^{(l)}$ corresponding to (k_1, k_2) and (k_2, k_1) with $k_1 < k_2$, obtaining new matrices $\mathbf{C}^{(l)}$. It is easy to see that, for every l, since $\mathbf{C}^{(l)}$ is obtained from $\tilde{\mathbf{C}}^{(l)}$ by adding columns together and since each row of $\tilde{\mathbf{C}}^{(l)}$ has exactly one unit element with the rest of them being zeros, $\mathbf{C}^{(l)}$ is again a clustering matrix of size $[n(n-1)/2] \times [m(m+1)/2]$. In particular, for indices i and k corresponding to nodes (i_1, i_2) and classes $(\Omega_{k_1}, \Omega_{k_2})$ with $i_1 < i_2$ and $k_1 \leq k_2$, one has $\mathbf{C}_{i,k}^{(l)} = 1$ if $i_1 \in \Omega_{k_1}$ and $i_2 \in \Omega_{k_2}$ or $i_1 \in \Omega_{k_2}$ and $i_2 \in \Omega_{k_1}$; $\mathbf{C}_{i,k}^{(l)} = 0$ otherwise. The process of vectorization of the model and removing redundancy is presented in Figure 1.

Using $\mathbf{C}^{(l)}$ and $\mathbf{q}^{(l)}$, one can rewrite equations (6) as

$$\mathbf{a}^{(l)} = \boldsymbol{\theta}^{(l)} + \boldsymbol{\xi}^{(l)} \qquad \text{with } \boldsymbol{\theta}^{(l)} = \mathbf{C}^{(l)}\mathbf{q}^{(l)}, \qquad l = 1, \ldots, L, \tag{7}$$

where $\mathbf{C}^{(l)} \in \mathcal{M}(M, N)$, $\boldsymbol{\theta}^{(l)} \in \mathbb{R}^N$, $\mathbf{q}^{(l)} \in \mathbb{R}^M$, $N = n(n-1)/2$ and $M = m(m+1)/2$. Here, for every i and l, components $\mathbf{a}_i^{(l)}$ of vector $\mathbf{a}^{(l)}$ are independent Bernoulli variables with $\mathbb{P}(\mathbf{a}_i^{(l)} = 1) = \boldsymbol{\theta}_i^{(l)}$, so that components of vectors $\boldsymbol{\xi}^{(l)}$ are also independent for different values of i or l.

If we had the time-independent SBM ($L = 1$) and the clustering matrix were known, equation (7) would reduce estimation of $\mathbf{q}^{(1)}$ to the linear regression prob-

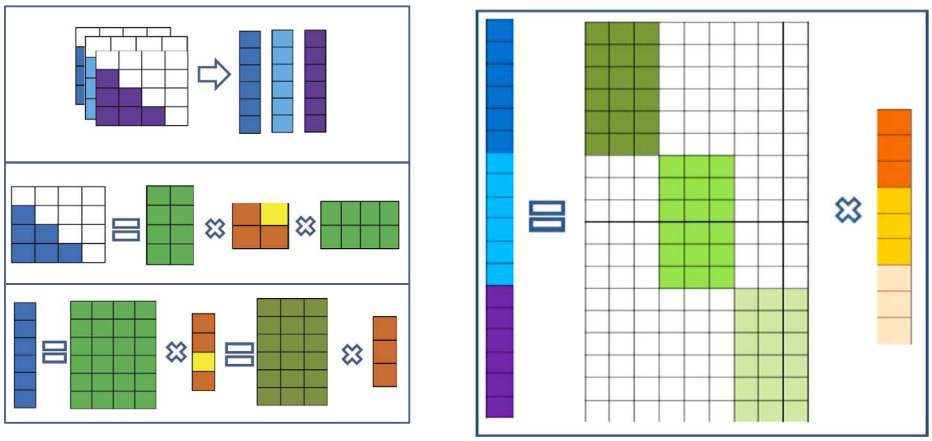

FIG. 1. *Vectorization of the probability tensor* $\boldsymbol{\Lambda}$ *with* $n = 4$, $m = 2$, $N = n(n-1)/2 = 6$, $M = m(m+1)/2 = 3$ *and* $L = 3$. *Left panel, top: transforming* $\boldsymbol{\Lambda}_{*,*,l}$ *into* $\boldsymbol{\theta}^{(l)}$, $l = 1, 2, 3$. *Left panel, middle:* $\boldsymbol{\Lambda}_{*,*,1} = \tilde{\mathbf{Z}}^{(1)}\mathbf{G}_{*,*,1}(\tilde{\mathbf{Z}}^{(1)})^T$. *Left panel, bottom:* $\boldsymbol{\theta}^{(1)} = \tilde{\mathbf{C}}^{(1)}\mathbf{g}^{(1)} = \mathbf{C}^{(1)}\mathbf{q}^{(1)}$. *In the left panel, redundant elements of* $\boldsymbol{\Lambda}$ *are white, redundant elements of* $\mathbf{G}$ *are yellow. Right panel:* $\boldsymbol{\theta} = \mathbf{C}\mathbf{q}$.

lem with independent sub-Gaussian (Bernoulli) errors. Since in the case of the DSBM, for each i, the elements $\mathbf{g}_i^{(l)}$, $l = 1, \dots, L$, of vector $\mathbf{g}_i$ represent the values of a smooth function, we combine vectors in (7) into matrices. Specifically, we consider matrices $\mathbf{A}, \mathbf{\Theta}, \mathbf{\Xi} \in \mathbb{R}^{N \times L}$ and $\mathbf{Q} \in \mathbb{R}^{M \times L}$ with columns $\mathbf{a}^{(l)}$, $\boldsymbol{\theta}^{(l)}$, $\boldsymbol{\xi}^{(l)}$ and $\mathbf{q}^{(l)}$, respectively. Note that if the group memberships of the nodes were constant in time, so that $\mathbf{C}^{(l)} \in \{0,1\}^{N \times M}$ were independent of l, formula (7) would imply

$$\mathbf{A} = \mathbf{\Theta} + \mathbf{\Xi}, \qquad \mathbf{\Theta} = \mathbf{ZQ} \qquad \text{if } \mathbf{C}^{(l)} = \mathbf{Z}, l = 1, \dots, L. \tag{8}$$

However, we consider the situations where nodes can switch group memberships in time and (8) is not true.

For this reason, we proceed with further vectorization. We denote $\mathbf{a} = \operatorname{vec}(\mathbf{A})$, $\boldsymbol{\theta} = \operatorname{vec}(\mathbf{\Theta})$ and $\mathbf{q} = \operatorname{vec}(\mathbf{Q})$ and observe that vectors $\mathbf{a}, \boldsymbol{\theta} \in \mathbb{R}^{NL}$ and $\mathbf{q} \in \mathbb{R}^{ML}$ are obtained by stacking vectors $\mathbf{a}^{(l)}$, $\boldsymbol{\theta}^{(l)}$ and $\mathbf{q}^{(l)}$ in (7) vertically for $l = 1, \dots, L$. Define a block diagonal matrix $\mathbf{C} \in \{0,1\}^{NL \times ML}$ with blocks $\mathbf{C}^{(l)}$, $l = 1, \dots, L$, on the diagonal. Then (7) implies that

$$\mathbf{a} = \boldsymbol{\theta} + \boldsymbol{\xi} \qquad \text{with } \boldsymbol{\theta} = \mathbf{Cq} = \mathbf{C} \operatorname{vec}(\mathbf{Q}), \tag{9}$$

where $\mathbf{a}_i$ are independent Bernoulli($\boldsymbol{\theta}_i$) variables, $i = 1, \dots, NL$.

Observe that if the matrix $\mathbf{C}$ were known, then equations in (9) would represent a regression model with independent Bernoulli errors. Moreover, matrix $\mathbf{C}^T\mathbf{C}$ is diagonal since matrices $(\mathbf{C}^{(l)})^T\mathbf{C}^{(l)} = (\mathbf{S}^{(l)})^2, l = 1, \dots, L$, are diagonal with $\mathbf{S}^{(l)}_{k_1,k_2} = \sqrt{N^{(l)}_{k_1,k_2}}$, where $N^{(l)}_{k_1,k_2}$ is the number of pairs (i_1, i_2) of nodes such that $i_1 < i_2$ and one node is in class Ω_{k_1} while another is in class Ω_{k_2} at time instant t_l:

$$N^{(l)}_{k_1,k_2} = \begin{cases} n^{(l)}_{k_1} n^{(l)}_{k_2} & \text{if } k_1 \neq k_2; \\ n^{(l)}_{k_1}(n^{(l)}_{k_1} - 1) & \text{if } k_1 = k_2. \end{cases} \tag{10}$$

REMARK 1 (Directed graph). Similar vectorization algorithm can be used when the dynamic network is constructed from directed graphs or graphs with self-loops. In the former case, the only redundant entries of matrices $\mathbf{\Lambda}_{*,*,l}$ would be the diagonal ones while, in the latter case, $\mathbf{\Lambda}$ has no redundant elements and no row removal is necessary.

REMARK 2 (Biclustering structures). Vectorization presented above can significantly simplify the inference in the so called bi-clustering models considered, for example, by Lee et al. (2010) and Gao et al. (2016). In those models, one needs to recover matrix $\mathbf{X}$ from observations of matrix $\mathbf{Y}$ given by $\mathbf{Y} = \mathbf{U}_1\mathbf{X}\mathbf{U}_2 + \mathbf{\Xi}$ where matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ are known and matrix $\mathbf{\Xi}$ has independent zero-mean Gaussian or sub-Gaussian entries. As long as there are no structural assumptions on matrix $\mathbf{X}$ (such as, e.g., low rank), one can apply vectorization and reduce the problem to the familiar nonparametric regression problem of the form $\mathbf{y} = \mathbf{Ux} + \boldsymbol{\xi}$ where matrix $\mathbf{U} = \mathbf{U}_1 \otimes \mathbf{U}_2$ is known, $\boldsymbol{\xi} = \operatorname{vec}(\mathbf{\Xi})$ is the vector with independent components and one needs to recover $\mathbf{x} = \operatorname{vec}(\mathbf{X})$ from observations $\mathbf{y} = \operatorname{vec}(\mathbf{Y})$.

3. Assumptions and estimation for the DSBM. It is reasonable to assume that the values of the probabilities $\mathbf{q}^{(l)}$ of connections do not change dramatically from one time instant to another. Specifically, we assume that for various $k = 1, \dots, M$, vectors $\mathbf{q}_k = (\mathbf{q}_k^{(1)}, \dots, \mathbf{q}_k^{(L)})$ represent values of some smooth functions, so that $\mathbf{q}_k^{(l)} = f_k(t_l)$, $l = 1, \dots, L$. In order to quantify this phenomenon, we assume that vectors $\mathbf{q}_k$ have sparse representation in some orthogonal basis $\mathbf{H} \in \mathbb{R}^{L \times L}$ with $\mathbf{H}^T\mathbf{H} = \mathbf{H}\mathbf{H}^T = \mathbf{I}_L$, so that vector $\mathbf{H}\mathbf{q}_k^T$ is sparse: it has only few large coefficients, the rest of the coefficients are small or equal to zero. This is a very common assumption in functional data analysis. For example, if $\mathbf{H}$ is the matrix of the Fourier transform and f_k belongs to a Sobolev space or $\mathbf{H}$ is a matrix of a wavelet transform and f_k belongs to a Besov space, the coefficients $\mathbf{H}\mathbf{q}_k^T$ of $\mathbf{q}_k^T$ decrease rapidly, and hence, vector $\mathbf{H}\mathbf{q}_k^T$ is sparse. In particular, one needs only few elements in vector $\mathbf{H}\mathbf{q}_k^T$ to represent $\mathbf{q}_k$ with high degree of accuracy. The extreme case occurs when the connection probabilities do not change in time, so that vector $\mathbf{q}_k$ has constant components: then, for the Fourier or a periodic wavelet transform, the vector $\mathbf{H}\mathbf{q}_k^T$ has only one nonzero element.

Denote $\mathbf{D} = \mathbf{Q}\mathbf{H}^T$ where matrix $\mathbf{Q}$ is defined in the previous section and $\mathbf{d} = \operatorname{vec}(\mathbf{D})$. Observe that vector $\mathbf{d}$ is obtained by stacking together the columns of matrix $\mathbf{D} = \mathbf{Q}\mathbf{H}^T$ while its transpose $\mathbf{D}^T = \mathbf{H}\mathbf{Q}^T$ has vectors $\mathbf{H}\mathbf{q}_k^T$ as its columns. Then sparsity of the matrix $\mathbf{D}$ can be controlled by imposing a complexity penalty $\|\mathbf{d}\|_0 = \|\mathbf{D}\|_0 = \|\mathbf{D}^T\|_0$ on matrix $\mathbf{D}$. Note that complexity penalty does not require the actual matrix $\mathbf{D}$ to have only few nonzero elements, it merely forces the procedure to keep only few large elements in $\mathbf{D}$ while setting the rest of the elements to zero, and hence, acts as a kind of hard thresholding. Note that by Theorem 1.2.22 of Gupta and Nagar (2000), one has

$$\mathbf{d} = \operatorname{vec}\bigl(\mathbf{Q}\mathbf{H}^T\bigr) = (\mathbf{H} \otimes \mathbf{I}_M)\operatorname{vec}(\mathbf{Q}) = (\mathbf{H} \otimes \mathbf{I}_M)\mathbf{q} = \mathbf{W}\mathbf{q}, \tag{11}$$

where $\mathbf{W} = (\mathbf{H} \otimes \mathbf{I}_M)$ is an orthogonal matrix such that $\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{I}_{ML}$. Denote

$$J \equiv J_M = \{j : \mathbf{d}_j \neq 0\}, \qquad \mathbf{d}_{J^C} = \mathbf{0}, \tag{12}$$

so that J is the set of indices corresponding to nonzero elements of the vector $\mathbf{d}$.

Consider a set of clustering matrices $\mathcal{C}(m, n, L)$ satisfying (2). At this point, we impose very mild assumption on $\mathcal{C}(m, n, L)$:

$$\log\bigl(\bigl|\mathcal{C}(m, n, L)\bigr|\bigr) \geq 2\log m. \tag{13}$$

Assumption (13) is only used for simplifying expression for the penalty. Indeed, until now, we allowed any collection of clustering matrices, so potentially, we can work with the case where all cluster memberships are fixed in advance (although this would be a totally trivial case). Condition (13) merely means that at least two nodes at some point in time can be assigned arbitrarily to any of m classes. Later, we shall consider some special cases such as fixed membership (no membership

switches over time) or limited change (only at most n_0 nodes can change their memberships between two consecutive time points).

We find m, J, $\mathbf{d}$ and $\mathbf{C}$ as a solution of the following penalized least squares optimization problem:

$$(\widehat{m}, \widehat{J}, \widehat{\mathbf{d}}, \widehat{\mathbf{C}}) \in \operatorname*{argmin}_{m,J,\mathbf{d},\mathbf{C}} \bigl[\|\mathbf{a} - \mathbf{C}\mathbf{W}^T\mathbf{d}\|^2 + \mathrm{Pen}(|J|, m)\bigr] \qquad \text{s.t. } \mathbf{d}_{J^c} = \mathbf{0}, \tag{14}$$

where $\mathbf{C} \in \mathcal{C}(m, n, L)$, $\mathbf{a}$ is defined in (9), $\mathbf{d} \in \mathbb{R}^{ML}$, $\mathbf{W} \in \mathbb{R}^{ML \times ML}$, $M = m(m+1)/2$ and

$$\mathrm{Pen}(|J|, m) = 11 \log\bigl(|\mathcal{C}(m, n, L)|\bigr) + \frac{11}{2}|J| \log\Bigl(\frac{25m^2L}{|J|}\Bigr). \tag{15}$$

Observe that the penalty in (15) consists of two parts. The first part accounts for the complexity of clustering and, therefore, allows one to obtain an estimator adaptive to the number of unknown groups m as long as the we can express the complexity of clustering in terms of m, n and L. The second term represents the price of estimating $|J|$ elements of vector $\mathbf{d}$ and finding those $|J|$ elements in this vector of length $m(m+1)L/2$.

Note that since minimization is carried out also with respect to m, optimization problem (14) should be solved separately for every $m = 1, \dots, n$, yielding $\widehat{\mathbf{d}}_M$, $\widehat{\mathbf{C}}_M$ and $\widehat{J}_M$. After that, one needs to select the value $\widehat{M} = \widehat{m}(\widehat{m}+1)/2$ that delivers the minimum in (14), so that

$$\widehat{\mathbf{d}} = \widehat{\mathbf{d}}_{\widehat{M}}, \qquad \widehat{\mathbf{C}} = \widehat{\mathbf{C}}_{\widehat{M}}, \qquad \widehat{J} = \widehat{J}_{\widehat{M}}. \tag{16}$$

Finally, due to (12), we set $\widehat{\mathbf{W}} = (\mathbf{H} \otimes \mathbf{I}_{\widehat{M}})$ and calculate

$$\widehat{\mathbf{q}} = \widehat{\mathbf{W}}^T\widehat{\mathbf{d}}, \qquad \widehat{\boldsymbol{\theta}} = \widehat{\mathbf{C}}\widehat{\mathbf{q}}. \tag{17}$$

We obtain $\widehat{\boldsymbol{\Lambda}}$ by packing vector $\widehat{\boldsymbol{\theta}}$ into the tensor and taking the symmetries into account.

4. Oracle inequalities for the DSBM. Denote the true value of tensor $\boldsymbol{\Lambda}$ by $\boldsymbol{\Lambda}^*$. Also, denote by m^* the true number of groups, by $\mathbf{q}^*$ and $\boldsymbol{\theta}^*$ the true values of $\mathbf{q}$ and $\boldsymbol{\theta}$ in (9) and by $\mathbf{C}^*$ the true value of $\mathbf{C}$. Denote by $\mathbf{D}^*$ and $\mathbf{d}^*$ the true values of matrix $\mathbf{D}$ and vector $\mathbf{d}$, respectively. Let $M^* = m^*(m^*+1)/2$ and $\mathbf{W}^* = (\mathbf{H} \otimes \mathbf{I}_{M^*})$ be true values of M and $\mathbf{W}$. Note that vector $\boldsymbol{\theta}^*$ is obtained by vectorizing $\boldsymbol{\Lambda}^*$ and then removing the redundant entries. Then it follows from (9) that

$$\mathbf{a} = \boldsymbol{\theta}^* + \boldsymbol{\xi} \qquad \text{with } \boldsymbol{\theta}^* = \mathbf{C}^*\mathbf{q}^* = \mathbf{C}^*\bigl(\mathbf{W}^*\bigr)^T\mathbf{d}^*. \tag{18}$$

Due to the relation between the ℓ_2 and the Frobenius norms, one has

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \le \|\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^*\|^2 \le 2\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2, \tag{19}$$

and the following statement holds.

THEOREM 1. *Consider a DSBM with a true matrix of probabilities* $\mathbf{\Lambda}^*$ *and the estimator* $\widehat{\mathbf{\Lambda}}$ *obtained according to* (14)–(17). *Let* $\mathcal{C}(m,n,L)$ *be a set of clustering matrices satisfying conditions* (2) *and* (13). *Then, for any* $t>0$, *with probability at least* $1-9e^{-t}$, *one has*

$$\frac{\|\widehat{\mathbf{\Lambda}}-\mathbf{\Lambda}^*\|^2}{n^2L} \le \min_{\substack{m,J,\mathbf{d}\\ \mathbf{C}\in\mathcal{C}(m,n,L)}} \left[\frac{6\|\mathbf{C}\mathbf{W}^T\mathbf{d}_{(J)}-\boldsymbol{\theta}^*\|^2}{n^2L}+\frac{4\,\mathrm{Pen}(|J|,m)}{n^2L}\right]+\frac{38t}{n^2L} \tag{20}$$

and

$$\mathbb{E}\left(\frac{\|\widehat{\mathbf{\Lambda}}-\mathbf{\Lambda}^*\|^2}{n^2L}\right) \le \min_{\substack{m,J,\mathbf{d}\\ \mathbf{C}\in\mathcal{C}(m,n,L)}} \left[\frac{6\|\mathbf{C}\mathbf{W}^T\mathbf{d}_{(J)}-\boldsymbol{\theta}^*\|^2}{n^2L}+\frac{4\,\mathrm{Pen}(|J|,m)}{n^2L}+\frac{342}{n^2L}\right], \tag{21}$$

where $\mathbf{d}_{(J)}$ *is the modification of vector* $\mathbf{d}$ *where all elements* $\mathbf{d}_j$ *with* $j\notin J$ *are set to zero*.

The proof of Theorem 1 is given in the Supplementary Material [Pensky (2019)]. Here, we just explain its idea. Note that if the values of m and C are fixed, the problem (14) reduces to a regression problem with a complexity penalty $\mathrm{Pen}(|J|,m)$. Moreover, if J is known, the optimal estimator $\widehat{\mathbf{d}}$ of $\mathbf{d}^*$ is just a projection estimator. Indeed, denote $\mathbf{\Upsilon}_{\mathbf{C}}=\mathbf{C}\mathbf{W}^T$ and let $\mathbf{\Upsilon}_{\mathbf{C},J}=(\mathbf{C}\mathbf{W}^T)_J$ be the reduction of matrix $\mathbf{C}\mathbf{W}^T$ to columns $j\in J$. Given $\widehat{m}$, $\widehat{J}$ and $\widehat{\mathbf{C}}$, one obtains $\widehat{M}=\widehat{m}(\widehat{m}+1)/2$, $\widehat{\mathbf{W}}=(\mathbf{H}\otimes\mathbf{I}_{\widehat{M}})$, $\mathbf{\Upsilon}_{\mathbf{C},J}=(\mathbf{C}\mathbf{W}^T)_J$ and $\widehat{\mathbf{\Upsilon}}_{\widehat{\mathbf{C}},\widehat{J}}=(\widehat{\mathbf{C}}\widehat{\mathbf{W}}^T)_{\widehat{J}}$. Let

$$\mathbf{\Pi}_{\mathbf{C},J}=\mathbf{\Upsilon}_{\mathbf{C},J}(\mathbf{\Upsilon}_{\mathbf{C},J}^T\mathbf{\Upsilon}_{\mathbf{C},J})^{-1}\mathbf{\Upsilon}_{\mathbf{C},J}^T, \qquad \widehat{\mathbf{\Pi}}_{\widehat{\mathbf{C}},\widehat{J}}=\widehat{\mathbf{\Upsilon}}_{\widehat{\mathbf{C}},\widehat{J}}(\widehat{\mathbf{\Upsilon}}_{\widehat{\mathbf{C}},\widehat{J}}^T\widehat{\mathbf{\Upsilon}}_{\widehat{\mathbf{C}},\widehat{J}})^{-1}\widehat{\mathbf{\Upsilon}}_{\widehat{\mathbf{C}},\widehat{J}}^T \tag{22}$$

be the projection matrices on the column spaces of $\mathbf{\Upsilon}_{\mathbf{C},J}$ and $\widehat{\mathbf{\Upsilon}}_{\widehat{\mathbf{C}},\widehat{J}}$, respectively. Then it is easy to see that $\widehat{\mathbf{\Upsilon}}_{\widehat{\mathbf{C}},\widehat{J}}\widehat{\mathbf{d}}=\widehat{\mathbf{\Pi}}_{\widehat{\mathbf{C}},\widehat{J}}\mathbf{a}$ and vector $\widehat{\mathbf{d}}$ is of the form

$$\widehat{\mathbf{d}}=(\widehat{\mathbf{\Upsilon}}_{\widehat{\mathbf{C}},\widehat{J}}^T\widehat{\mathbf{\Upsilon}}_{\widehat{\mathbf{C}},\widehat{J}})^{-1}\widehat{\mathbf{\Upsilon}}_{\widehat{\mathbf{C}},\widehat{J}}^T\mathbf{a}. \tag{23}$$

Hence, the values of $\widehat{m}$, $\widehat{J}$ and $\widehat{\mathbf{C}}$ can be obtained as a solution of the following optimization problem:

$$(\widehat{\mathbf{C}},\widehat{m},\widehat{J})\in\operatorname*{argmin}_{m,J,\mathbf{C}}\big[\|\mathbf{a}-\mathbf{\Pi}_{\mathbf{C},J}\mathbf{a}\|^2+\mathrm{Pen}\big(|J|,m\big)\big] \qquad \text{s.t. } \mathbf{C}\in\mathcal{C}(m,n,L),$$

where $\mathbf{\Pi}_{\mathbf{C},J}$ and $\mathrm{Pen}(|J|,m)$ are defined in (22) and (15), respectively. After that, we use the arguments that are relatively standard in the proofs of oracle inequalities for the penalized least squares estimators.

Note that $\|\mathbf{C}\mathbf{W}^T\mathbf{d}_{(J)}-\boldsymbol{\theta}^*\|^2$ in the right-hand sides of expressions (20) and (21), is the bias term that quantifies how well one can estimate the true values

of probabilities $\boldsymbol{\theta}^*$ by blocking them together, averaging the values in each block and simultaneously setting all but $|J|$ elements of vector $\mathbf{d}$ to zero. If $|J|$ is too small, then $\mathbf{d}$ will not be well represented by its truncated version $\mathbf{d}_{(J)}$ and the bias will be large. The penalty represents the stochastic error and constitutes the "price" for choosing too many blocks and coefficients. In particular, the second term $(11/2)|J|\log(25\ m^2L/|J|)$ in (15) is due to the need of finding and estimating $|J|$ elements of the $Lm(m+1)/2$-dimensional vector. The first term, $\log(|\mathcal{C}(m,n,L)|)$, accounts for the difficulty of clustering and is due to application of the union bound in probability.

Theorem 1 holds for any collection $\mathcal{C}(m,n,L)$ of clustering matrices satisfying assumption (13). In order to obtain some specific results, denote by $\mathcal{Z}(m,n,n_0,L)$ the collection of clustering matrices corresponding to the situation where at most n_0 nodes can change their memberships between any two consecutive time points, so that

$$|\mathcal{Z}(m,n,n_0,L)| = m^n\left[\binom{n}{n_0}m^{n_0}\right]^{L-1}, \tag{24}$$

yielding $|\mathcal{Z}(m,n,0,L)| = m^n$ and $|\mathcal{Z}(m,n,n,L)| = m^{nL}$. Note that the case of $n_0 = 0$ corresponds to the scenario where the group memberships of the nodes are constant and do not depend on time while the case of $n_0 = n$ means that memberships of all nodes can change arbitrarily from one time instant to another. Since

$$\log\left[\binom{n}{n_0}m^{n_0}\right] \le n_0\log\left(\frac{mne}{n_0}\right),$$

formulae (15) and (24) immediately yield the following corollary.

COROLLARY 1. *Consider a DSBM with a true matrix of probabilities* $\boldsymbol{\Lambda}^*$ *and estimator* $\widehat{\boldsymbol{\Lambda}}$ *obtained according to* (14)–(17) *where* $\mathcal{C}(m,n,L) = \mathcal{Z}(m,n,n_0,L)$. *Then inequalities* (20) *and* (21) *hold with*

$$\operatorname{Pen}(|J|,m) = 11\left[n\log m + n_0(L-1)\log\left(\frac{mne}{n_0}\right) + \frac{|J|}{2}\log\left(\frac{25\ m^2L}{|J|}\right)\right]. \tag{25}$$

It is easy to see that the first term in (25) accounts for the uncertainty of the initial clustering, the second term is due to the changes in the group memberships of the nodes over time (indeed, if $n_0 = 0$, this term just vanishes) while the last term is identical to the second term in the expression for the generic penalty (15). While we elaborate only on the special case where the collection of clustering matrices is given by (24), one can easily produce results similar to Corollary 1 for virtually any nodes' memberships scenario.

REMARK 3 (The SBM). Theorem 1 provides an oracle inequality in the case of a time-independent SBM ($L = 1$). Indeed, in this case, by taking $\mathbf{H} = 1$ and $\mathbf{W} = \mathbf{I}_M$, obtain for any $t > 0$

$$\frac{\mathbb{E}\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}^*\|^2}{n^2} \le \min_{\substack{m,J,\mathbf{q} \\ \mathbf{C}\in\mathcal{M}(m,n)}} \left[\frac{6\|\mathbf{C}\mathbf{q}_{(J)} - \boldsymbol{\theta}^*\|^2}{n^2} + \frac{44\log m}{n} + \frac{22|J|}{n^2}\log\left(\frac{25\,m^2}{|J|}\right)\right] + \frac{342}{n^2} \tag{26}$$

and a similar result holds for the probability. Note that if $|J| = m(m+1)/2$, our result coincides with the one of Gao, Lu and Zhou (2015). However, if many groups have zero probability of connection, then $|J|$ is small and the right-hand side of (26) can be asymptotically smaller than $n^{-1}\log m + n^{-2}m^2$ obtained in Gao, Lu and Zhou (2015). In addition, our oracle inequality is non-asymptotic and the estimator is naturally adaptive to the unknown number of classes. [Gao et al. (2016) obtained adaptive estimators but not via an oracle inequality.]

Corollary 1 quantifies the stochastic error term in Theorem 1. The size of the bias depends on the level of sparsity of coefficients of functions $\mathbf{q}_k$ in the basis $\mathbf{H}$ and on the constitution of classes. While one can study a variety of scenarios, in order to be specific, we consider the case of a *balanced network model* where the sizes of all the classes are proportional to each other, in particular, for some absolute constants $0 < \aleph_1 \le 1 \le \aleph_2 < \infty$, one has

$$\aleph_1\frac{n}{m} \le n_k^{(l)} \le \aleph_2\frac{n}{m}, \qquad k = 1,\dots,m, l = 1,\dots,L, \tag{27}$$

where $n_k^{(l)}$ the number of nodes in class k at the moment t_l.

Note that the condition (27) is very common in studying random network models [see, e.g., Gao et al. (2017) or Amini and Levina (2018) among others]. In addition, if class memberships are generated from the multinomial distribution with the vector of probabilities $(\pi_1,\dots,\pi_m)$, and $C_1/m \le \pi_i \le C_2/m$ for some constants $0 < C_1 < C_2 < \infty$, as it is done in, for example, Bickel and Chen (2009), condition (27) holds with high probability.

In particular, we consider networks that satisfy condition (27) but yet allow only n_0 nodes switch their memberships between time instances. We denote the corresponding set of clustering matrices by $\mathcal{Z}_{\text{bal}}(m,n,n_0,L,\aleph_1,\aleph_2)$. It would seem that condition (27) should make clustering much simpler. However, as Lemma 1 below shows, this reduction does not makes estimation significantly easier since the complexity of the set of balanced clustering matrices $\log|\mathcal{Z}_{\text{bal}}(m,n,n_0,L,\aleph_1,\aleph_2)|$ is smaller than the complexity of the set of unrestricted clustering matrices $\log|\mathcal{Z}(m,n,n_0,L)|$ only by, at most, a constant factor.

LEMMA 1 (Balanced network model complexity). *If* $n \geq \sqrt{en_0^3}$, *then*

$$\log|\mathcal{Z}_{\mathrm{bal}}(m,n,n_0,L,\aleph_1,\aleph_2)| \geq \frac{1}{4}\biggl[n\log m + (L-1)n_0\log\biggl(\frac{mne}{n_0}\biggr)\biggr]. \tag{28}$$

Then one can use the same penalty that was considered in Corollary 1, so that Theorem 1 yields the following result.

THEOREM 2. *Consider a balanced DSBM satisfying condition* (27). *Let* $\mathbf{\Lambda}^*$ *be the true matrix of probabilities*, m^* *be the true number of classes*, $M^* = m^*(m^*+1)/2$, $\mathbf{Q}^*$ *be the true matrix of probabilities of connections for pairs of classes and* $\mathbf{D}^* = \mathbf{Q}^*\mathbf{H}$. *If* $n \geq \sqrt{en_0^3}$ *and the estimator* $\widehat{\mathbf{\Lambda}}$ *is obtained as a solution of optimization problem* (14) *with the penalty* (25) *where*

$$J = \bigcup_{k=1}^{M} J_k, \tag{29}$$

then, for any $t > 0$, *with probability at least* $1 - 9e^{-t}$, *one has*

$$\frac{\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}^*\|^2}{n^2L} \leq \min_J\biggl\{\frac{6\aleph_2^2}{(m^*)^2L}\sum_{k=1}^{M^*}\sum_{l\notin J_k}\bigl(\mathbf{D}^*_{k,l}\bigr)^2 + \frac{4\,\mathrm{Pen}(|J|,m^*)}{n^2L}\biggr\} + \frac{38t}{n^2L} \tag{30}$$

and a similar result holds for the expectation.

In order to obtain specific upper bounds in (30), we need to impose some assumptions on the smoothness of functions $\mathbf{Q}^*_{k,*}$, $k = 1,\dots,M^*$. For the sake of brevity, we assume that all vectors $\mathbf{D}^*_{k,*}$, $k = 1,\dots,M^*$, behave similarly with respect to the basis $\mathbf{H}$ (generalization to the case where this is not true is rather pedestrian but very cumbersome as we point out in Section 8, Discussion).

(A0). There exist absolute constants ν_0 and K_0 such that

$$\sum_{l=1}^{L}(l-1)^{2\nu_0}\bigl(\mathbf{D}^*_{k,l}\bigr)^2 \leq K_0, \qquad k = 1,\dots,M^*. \tag{31}$$

COROLLARY 2. *Let the conditions of Theorem* 2 *hold and* $\mathbf{D}^*_{k,*}$ *satisfy assumption* (31). *If the estimator* $\widehat{\mathbf{\Lambda}}$ *is obtained as a solution of optimization problem* (14), *then for any* $t > 0$, *with probability at least* $1 - 9e^{-t}$, *one has*

$$\begin{aligned}\frac{\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}^*\|^2}{n^2L} \leq \tilde{K}_0\biggl(&\min\biggl\{\frac{1}{L}\biggl[\biggl(\frac{m^*}{n}\biggr)^2\log\biggl(\frac{n}{m^*}\biggr)\biggr]^{\frac{2\nu_0}{2\nu_0+1}}, \biggl(\frac{m^*}{n}\biggr)^2\biggr\}\\ &+ \frac{\log m^*}{nL} + \frac{n_0}{n^2}\log\biggl(\frac{m^*ne}{n_0}\biggr) + \frac{t}{n^2L}\biggr)\end{aligned} \tag{32}$$

and a similar result holds for the expectation. Here, $\tilde{K}_0$ *is an absolute constant that depends on* K_0, ν_0, $\aleph_1$ *and* $\aleph_2$ *only.*

5. The lower bounds for the risk for the DSBM. In order to prove that the estimator obtained as a solution of optimization problem (14) is minimax optimal, we need to show that the upper bounds in Corollaries 1 and 2 coincide with the minimax lower bounds obtained under similar constraints. For the sake of derivation of lower bounds for the error, we impose mild conditions on the orthogonal matrix $\mathbf{H}$ as follows: for any binary vector $\boldsymbol{\omega} \in \{0,1\}^L$ one has

$$\|\mathbf{H}^T\boldsymbol{\omega}\|_\infty \le \|\boldsymbol{\omega}\|_1/\sqrt{L} \quad \text{and} \quad \mathbf{H}\mathbf{1} = \sqrt{L}\mathbf{e}_1, \tag{33}$$

where $\mathbf{1} = (1,1,\dots,1)^T$ and $\mathbf{e}_1 = (1,0,\dots,0)^T$. Assumptions (33) are not restrictive. In fact, they are satisfied for a variety of common orthogonal transforms such as the Fourier transform or a periodic wavelet transforms.

First, we derive the lower bounds for the risk under the assumption that vector $\mathbf{d}$ is l_0-sparse and has only s nonzero components. Let $\mathcal{G}_{m,L,s}$ be a collection of tensors such that $\mathbf{G} \in \mathcal{G}_{m,L,s}$ implies that the vectorized versions $\mathbf{q}$ of $\mathbf{G}$ can be written as $\mathbf{q} = \mathbf{W}^T\mathbf{d}$ with $\|\mathbf{d}\|_0 \le s$. In order to be more specific, we consider the collection of clustering matrices $\mathcal{Z}(m,n,n_0,L)$ with cardinality given by (24) that corresponds to the situation where at most n_0 nodes can change their memberships between consecutive time instants. In this case, $\mathrm{Pen}(|J|,m)$ is defined in (25).

THEOREM 3. *Let orthogonal matrix $\mathbf{H}$ satisfy condition (33). Consider the DSBM where $\mathbf{G} \in \mathcal{G}_{m,L,s}$ with $s \ge \kappa m^2$ where $\kappa > 0$ is independent of m, n and L. Denote $\gamma = \min(\kappa, 1/2)$ and assume that $L \ge 2$, $n \ge 2m$, $n_0 \le \min(\gamma n, 4/3\gamma n m^{-1/9})$ and s is such that*

$$s^2 \log(2LM/s) \le 68LMn^2. \tag{34}$$

Then

$$\inf_{\widehat{\boldsymbol{\Lambda}}} \sup_{\substack{\mathbf{G}\in\mathcal{G}_{m,L,s} \\ \mathbf{C}\in\mathcal{Z}(m,n,n_0,L)}} \mathbb{P}_{\boldsymbol{\Lambda}}\left\{ \frac{\|\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\|^2}{n^2L} \ge C(\gamma)\left(\frac{\log m}{nL} + \frac{n_0}{n^2}\log\left(\frac{mne}{n_0}\right) + \frac{s\log(Lm^2/s)}{n^2L}\right)\right\} \ge \frac{1}{4}, \tag{35}$$

where $\widehat{\boldsymbol{\Lambda}}$ is any estimator of $\boldsymbol{\Lambda}$, $\mathbb{P}_{\boldsymbol{\Lambda}}$ is the probability under the true value of the tensor $\boldsymbol{\Lambda}$ and $C(\gamma)$ is an absolute constant that depends on γ only.

Theorem 3 ensures that if vector $\mathbf{d}$ has only s nonzero components, then the upper bounds in Corollary 1 are optimal up to a constant. In order to provide a similar assertion in the case of Corollary 2, we assume that rows of matrix $\mathbf{D}$ are l_2-sparse. For this purpose, we consider a collection of tensors $\mathcal{G}_{m,L,v_0}$ such that $\mathbf{G} \in \mathcal{G}_{m,L,v_0}$ implies that $\mathbf{Q} = \mathbf{D}\mathbf{H}$ and rows $\mathbf{D}_{k,*}$ of matrix $\mathbf{D}$ satisfy condition (31). Let as before $\mathcal{Z}_{\mathrm{bal}}(m,n,n_0,L,\aleph_1,\aleph_2)$ be a collection of clustering matrices satisfying condition (27) and such that at most n_0 nodes change their memberships

between two consecutive time instances. The following statement ensures that the upper bounds in Corollary 2 are minimax optimal up to a constant factor.

THEOREM 4. *Let orthogonal matrix* $\mathbf{H}$ *satisfy condition* (33). *Consider the DSBM where* $\mathbf{G} \in \mathcal{G}_{m,L,\nu_0}$ *with* $\nu_0 > 1/2$, $L \geq 2$ *and* $n \geq 2m$. *Then, for any absolute constants* $0 < \aleph_1 \leq 1 \leq \aleph_2 < \infty$, *one has*

$$\inf_{\widehat{\mathbf{\Lambda}}} \sup_{\substack{\mathbf{G}\in\mathcal{G}_{m,L,s} \\ \mathbf{C}\in\mathcal{Z}_{\mathrm{bal}}}} \mathbb{P}_{\mathbf{\Lambda}}\left\{\frac{\|\widehat{\mathbf{\Lambda}}-\mathbf{\Lambda}\|^2}{n^2 L} \geq C\left[\min\left\{\frac{1}{L}\left[\left(\frac{m}{n}\right)^2\right]^{\frac{2\nu_0}{2\nu_0+1}}; \left(\frac{m}{n}\right)^2\right\} + \frac{\log m}{nL} + \frac{n_0}{n^2}\log\left(\frac{mne}{n_0}\right)\right]\right\} \geq \frac{1}{4}, \tag{36}$$

where $\mathcal{Z}_{\mathrm{bal}}$ *stands for* $\mathcal{Z}_{\mathrm{bal}}(m, n, n_0, L, \aleph_1, \aleph_2)$, $\widehat{\mathbf{\Lambda}}$ *is any estimator of* $\mathbf{\Lambda}$, $\mathbb{P}_{\mathbf{\Lambda}}$ *is the probability under the true value of the tensor* $\mathbf{\Lambda}$ *and* C *is an absolute constant independent of* n, m *and* L.

Theorems 3 and 4 confirm that the estimator constructed above is *minimax optimal up to a constant* if $\mathbf{G} \in \mathcal{G}_{m,L,s}$ and $\mathbf{C} \in \mathcal{Z}(m, n, n_0, L)$, or $\mathbf{G} \in \mathcal{G}_{m,L,\nu_0}$ and $\mathbf{C} \in \mathcal{Z}_{\mathrm{bal}}(m, n, n_0, L, \aleph_1, \aleph_2)$.

Note that the terms $\log m/(nL)$ and $n_0 n^{-2}\log(mne/n_0)$ in (35) and (36) correspond to, respectively, the error of initial clustering and the clustering error due to membership changes. The remaining terms are due to nonparametric estimation and model selection. Assumptions (33) and (34) are purely technical and are necessary to ensure that the "worst case scenario" tensor $\mathbf{G}$ of connection probabilities has nonnegative components. As we mentioned earlier, conditions (33) are totally nonrestrictive. Condition (34) in Theorem 3 holds whenever representation of the tensor of probabilities in the basis $\mathbf{H}$ is at least somewhat sparse. Indeed, if there is absolutely no sparsity (which is a very implausible scenario when smooth functions are represented in a basis) and $s \approx ML$, then condition (34) reduces to $m(m+1)L \leq Cn^2$ and will still be true if L is relatively small. If L is large, the situation where $s \approx ML$ is very unlikely. Assumption that $s \geq \kappa m^2$ for some $\kappa > 0$ independent of m, n and L, restricts the sparsity level and ensures that one does not have too many classes where nodes have no interactions with each other or members of other classes.

Finally, it is also worth keeping in mind that all assumptions in Theorems 3 and 4 are used for the derivation of the minimax lower bounds for the risk and are not necessary for either the construction of the estimator $\widehat{\mathbf{\Lambda}}$ of $\mathbf{\Lambda}$ in (14) or for the assessment of its precision in Theorems 1 and 2.

6. The uniformly sparse DSBM. In the current literature, the notion of the sparse SBM refers to the case where the entries of the matrix of the connection probabilities are uniformly small: $\mathbf{\Lambda} = \rho_n \mathbf{\Lambda}^{(0)}$ with $\|\mathbf{\Lambda}^{(0)}\|_\infty = 1$ and $\rho_n \to 0$ as

$n \to \infty$. The concept is based on the idea that when the number of nodes in a network grow, the probabilities of connections between them decline. The minimax study of the sparse SBM has been carried out by Klopp, Tsybakov and Verzelen (2017). The logical generalization of the sparse SBM of this type would be the sparse DSBM where the elements of the tensor $\mathbf{\Lambda}$ are bounded above by ρ_n where $\rho_n \to 0$ as n grows. We refer to this kind of network as *uniformly sparse*.

On the other hand, not all networks become uniformly sparse as $n \to \infty$. Indeed, in the real world, when a network grows, the number of communities increases and, while the probabilities of connections for majority of pairs of groups become very small, some of the of pairs groups will still maintain high connection probabilities. We refer to this type of network as *nonuniformly sparse*. The idea of such a network has been elaborated in the recent paper of Borgs et al. (2016). The authors considered heavy-tailed sparse graphs such that, in the context of the SBM, one still has $\mathbf{\Lambda} = \rho_n \mathbf{\Lambda}^{(0)}$ but the elements of $\mathbf{\Lambda}^{(0)}$ are no longer bounded by one but by a quantity that grows with n.

While distinguishing between very small probabilities might be essential in a clustering problem, it is not so necessary in the problem of estimation of the tensor of the connection probabilities studied in the present paper. Indeed, it is a common knowledge that, in the nonparametric regression model, in order to obtain the best error rates, one needs to replace small elements of the vector of interest by zeros rather than estimating them. Similarly, if the network is nonuniformly sparse, that is, some pairs of groups have probabilities of connections equal or very close to zero, one would obtain an estimator with better overall precision by setting those very small connection probabilities to zeros. Although nowhere in the present paper we make an assumption that a network is sparse and, moreover, consideration of the nonuniformly sparse SBM or DSBM is not one our objectives, this paper naturally provides the tools for minimax optimal statistical estimation in such models that deliver results with very little additional work.

In addition, the techniques developed in this paper allow, with some additional work, to extend results obtained in Klopp, Tsybakov and Verzelen (2017) to the dynamic setting. However, majority of their results depend upon solution of optimization problem (14) under the restriction that $\|\mathbf{W}^T \mathbf{d}\|_\infty \leq \rho_n$ which requires representation of the estimator via a different projection operator and will result in more cumbersome calculations. Therefore, we avoid studying this new optimization problem and only extend Corollary 2.2 of Klopp, Tsybakov and Verzelen (2017) that handles the case of the balanced model without placing the above-mentioned restriction. For this purpose, consider a small ρ_n and denote

$$r_n(m) = \max\bigl(\rho_n, m^2/n^2\bigr). \tag{37}$$

Similar to (14), we find m, J, $\mathbf{d}$ and $\mathbf{C}$ as a solution of the following penalized

least squares optimization problem:

$$(\widehat{m}, \widehat{J}, \widehat{\mathbf{d}}, \widehat{\mathbf{C}}) \in \operatorname*{argmin}_{m, J, \mathbf{d}, \mathbf{C}} \big[\|\mathbf{a} - \mathbf{C}\mathbf{W}^T\mathbf{d}\|^2 + \lambda_0 r_n(m)\,\mathrm{Pen}\big(|J|, m\big)\big] \qquad \text{s.t. } \mathbf{d}_{J^c} = \mathbf{0} \tag{38}$$

where $\mathbf{C} \in \mathcal{Z}_{\mathrm{bal}}(m, n, n_0, L, \aleph_1, \aleph_2)$, $\mathbf{a}$ is defined in (9), $\mathbf{d} \in \mathbb{R}^{ML}$, $\mathbf{W} \in \mathbb{R}^{ML \times ML}$, $M = m(m+1)/2$, $\mathrm{Pen}(|J|, m)$ is defined in (25) and λ_0 is a tuning parameter that is bounded above and below by a constant.

In order the estimator has the uniform sparsity property, we need to make sure that transformation $\mathbf{H}$ is such that, whenever it is used for sparse representation of smooth functions, the maximum absolute value of the estimator obtained by truncation of the vector of coefficients is bounded above by a constant factor of the maximum absolute value of the original function. In particular, we denote the projection matrix on the column space of matrix $(\mathbf{C}\mathbf{W}^T)_J$ by $\mathbf{\Pi}_{\mathbf{C},J}$ and impose the following condition on the transformation matrix $\mathbf{H}$:

(A1). There exists an absolute constant B_0 such that for any $\mathbf{C} \in \mathcal{Z}_{\mathrm{bal}}(m, n, n_0, L, \aleph_1, \aleph_2)$ and any vector $\boldsymbol{\theta}$

$$\big\|\mathbf{\Pi}^{\perp}_{\mathbf{C},J}\boldsymbol{\theta}\big\|_\infty = \|\theta - \mathbf{\Pi}_{\mathbf{C},J}\theta\|_\infty \le B_0\|\theta\|_\infty. \tag{39}$$

Let, as before, $\mathbf{\Lambda}^*$ be the true matrix of probabilities, m^* be the true number of classes, $M^* = m^*(m^*+1)/2$, $\mathbf{C}^*$ be the true clustering matrix, $\mathbf{Q}^*$ be the true matrix of probabilities of connections for pairs of classes, $\mathbf{D}^* = \mathbf{Q}^*\mathbf{H}$, $\mathbf{d}^* = \mathrm{vec}(\mathbf{D}^*)$, $\boldsymbol{\theta}^* = \mathbf{C}^*(\mathbf{W}^*)^T\mathbf{d}^*$ and $\mathbf{W}^* = \mathbf{H} \otimes \mathbf{I}_{M^*}$.

THEOREM 5. *Consider a balanced DSBM satisfying condition* (27). *Let matrix* $\mathbf{H}$ *be such that condition* (39) *is satisfied and* $\|\mathbf{\Lambda}^*\|_\infty \le \rho_n^*$. *If* $\rho_n \ge \rho_n^*$, $n \ge \sqrt{en_0^3}$ *and the estimator* $\widehat{\mathbf{\Lambda}}$ *is obtained as a solution of optimization problem* (38), *then, for an absolute constant* $\tilde{C}_0$ *and any* $t > 0$, *with probability at least* $1 - 9e^{-t}$, *one has*

$$\frac{\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}^*\|^2}{n^2 L} \le \tilde{C}_0 \min_J \left\{ \frac{\|\mathbf{\Pi}^{\perp}_{\mathbf{C}^*,J}\boldsymbol{\theta}^*\|^2}{n^2 L} + \frac{r_n(m^*)[\mathrm{Pen}(|J|, m^*) + t]}{n^2 L} \right\}, \tag{40}$$

where $\mathbf{\Pi}_{\mathbf{C}^*,J}$ *is the projection matrix on the column space of* $(\mathbf{C}^*\mathbf{W}^{*T})_J$ *and* $\tilde{C}_0$ *is an absolute constant that depends on* B_0, $\aleph_1$ *and* $\aleph_2$ *only.*

In particular, if condition (31) *holds with* K_0 *replaced with* $\rho_n^* K_0$, *then*

$$\frac{\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}^*\|^2}{n^2 L} \le \tilde{K}_0 r_n(m^*) \bigg(\min\bigg\{ \frac{1}{L}\bigg[\bigg(\frac{m^*}{n}\bigg)^2 \log\bigg(\frac{n}{m^*}\bigg)\bigg]^{\frac{2\nu_0}{2\nu_0+1}}, \bigg(\frac{m^*}{n}\bigg)^2 \bigg\} + \frac{\log m^*}{nL} + \frac{n_0}{n^2}\log\bigg(\frac{m^* n e}{n_0}\bigg) + \frac{t}{n^2 L}\bigg). \tag{41}$$

Here, $\tilde{K}_0$ *is an absolute constant that depends on* B_0, K_0, ν_0, $\aleph_1$ *and* $\aleph_2$ *only. Results similar to* (40) *and* (41) *hold for the expectations.*

7. Dynamic graphon estimation. Consider the situation where tensor $\mathbf{\Lambda}$ is generated by a dynamic graphon f, so that $\mathbf{\Lambda}$ is given by expression (1) where function $f : [0,1]^3 \to [0,1]$ is such that $f(x,y,t) = f(y,x,t)$ for any t and $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)$ is a random vector sampled from a distribution $\mathbb{P}_\zeta$ supported on $[0,1]^n$.

Given an observed adjacency tensor $\mathbf{B}$ sampled according to model (1), the graphon function f is not identifiable since the topology of a network is invariant with respect to any change of labeling of its nodes. Therefore, for any f and any measure-preserving bijection $\mu : [0,1] \to [0,1]$ (with respect to Lebesgue measure), the functions $f(x,y,t)$ and $f(\mu(x),\mu(y),t)$ define the same probability distribution on random graphs. For this reason, we are considering equivalence classes of graphons. Note that in order it is possible to compare clustering of nodes across time instants, we introduce an assumption that there are no label switching in time, that is, every node carries the same label at any time t_l, so that the function μ is independent of t.

Under this condition, we further assume that probabilities $\mathbf{\Lambda}_{i,j,l}$ do not change drastically from one time point to another, that is, that, for every x and y, functions $f(x,y,t)$ are smooth in t. We shall also assume that f is piecewise smooth in x and y. In order to quantify those assumptions, for each $x, y \in [0,1]^2$, we consider a vector $\mathbf{f}(x,y) = (f(x,y,t_1), \dots, f(x,y,t_L))^T$ and an orthogonal transform $\mathbf{H}$ used in the previous sections. We assume that elements $\mathbf{v}_l(x,y)$ of vector $\mathbf{v}(x,y) = \mathbf{H}\mathbf{f}(x,y)$ satisfy the following assumption:

(A2). There exist constants $0 = \beta_0 < \beta_1 < \cdots < \beta_r = 1$ and $\nu_1, \nu_2, K_1, K_2 > 0$ such that for any $x, x' \in (\beta_{i-1}, \beta_i]$ and $y, y' \in (\beta_{j-1}, \beta_j]$, $1 \le i, j \le r$, one has

$$\big[\mathbf{v}_l(x,y) - \mathbf{v}_l(x',y')\big]^2 \le K_1\big[|x-x'| + |y-y'|\big]^{2\nu_1}, \tag{42}$$

$$\sum_{l=1}^{L} (l-1)^{2\nu_2} \mathbf{v}_l^2(x,y) \le K_2. \tag{43}$$

Note that, for a graphon corresponding to the DSBM model, on each of the rectangles $(\beta_{i-1}, \beta_i] \times (\beta_{j-1}, \beta_j]$, functions $\mathbf{v}_l(x,y)$ are constant, so that $\mathbf{v}_l(x,y) = 0$ for $l = 2, \dots, L$ and $\nu_1 = \infty$.

We denote the class of graphons satisfying assumptions (1), (42) and (43) by $\Sigma(\nu_1, \nu_2, K_1, K_2)$. In order to estimate the dynamic graphon, we approximate it by an appropriate DSBM and then estimate the probability tensor of the DSBM. Note that, since ν_1, ν_2, K_1 and K_2 in Assumption **A** are independent of x and y, one can simplify the optimization procedure in (14).

Let $\mathbf{Q}$ be the matrix defined in (8) and (9). Note that since random variables $\zeta_1, \dots, \zeta_n$ are time-independent, we can approximate the graphon by a DSBM where group memberships of the nodes do not change in time. Hence, matrices $\mathbf{C}^{(l)}$ are independent of l, so that $\mathbf{C}^{(l)} = \mathbf{Z}$, (8) holds and $\mathbf{\Theta} = \mathbf{ZQ}$. Denote $\mathbf{X} = \mathbf{AH}^T$. Denote by $\mathbf{V}$ and $\mathbf{\Phi}$ the matrices of the coefficients of $\mathbf{Q}$ and $\mathbf{\Theta}$ in the

transform $\mathbf{H}$: $\mathbf{V} = \mathbf{Q}\mathbf{H}^T$ and $\mathbf{\Phi} = \mathbf{\Theta}\mathbf{H}^T$. Then, by (8), $\mathbf{\Theta} = \mathbf{Z}\mathbf{V}\mathbf{H}$ and $\mathbf{\Phi} = \mathbf{Z}\mathbf{V}$. Note that each row of the matrices $\mathbf{V}$ and $\mathbf{\Phi}$ corresponds to one spatial location. Since, due to (43), the coefficients in the transform $\mathbf{H}$ decrease uniformly irrespective of the location, one can employ $L_1 < L$ columns instead of L columns in the final representations of $\mathbf{Q}$ and $\mathbf{\Theta}$. In order to simplify our presentation, we denote $L_1 = L^\rho$ where $0 < \rho \le 1$ and use the optimization procedure (14) to find m, ρ, $\mathbf{V}^{(\rho)}$ and $\mathbf{Z}$ where $\mathbf{V}^{(\rho)}$ is the submatrix of $\mathbf{V}$ with columns $\mathbf{V}_{*,j}$, $1 \le j \le L^\rho$. Due to $|J| = ML^\rho$ and $0.5\, m^2 L^\rho \le |J| \le m^2 L^\rho$, in this case optimization problem (14) can be reformulated as

$$
\begin{aligned}
(\widehat{m}, \widehat{\rho}, \widehat{\mathbf{V}}^{(\widehat{\rho})}, \widehat{\mathbf{Z}}) \in \operatorname*{argmin}_{\substack{m,\rho,\mathbf{V}^{(\rho)} \\ \mathbf{Z}\in\mathcal{Z}(m,n,0,L)}} \Big[&\|\mathbf{X}^{(\rho)} - \mathbf{Z}\mathbf{V}^{(\rho)}\|^2 \\
&+ 11n\log m + \frac{11}{2} m^2 L^\rho \log\bigl(25L^{1-\rho}\bigr)\Big],
\end{aligned} \tag{44}
$$

where $\mathcal{Z}(m, n, 0, L)$ is defined in (24). Then the estimation algorithm appears as follows:

1. Apply transform $\mathbf{H}$ to the data matrix $\mathbf{A}$ obtaining matrix $\mathbf{X} = \mathbf{A}\mathbf{H}^T$.
2. Consider a set $\Re = \{\rho \in [0, 1] : L^\rho \text{ is an integer}\}$. For every $\rho \in \Re$, remove all columns $\mathbf{X}_{*,l}$ with $l \ge L^\rho + 1$ obtaining matrix $\mathbf{X}^{(\rho)}$ with $\mathbb{E}\mathbf{X}^{(\rho)} = \mathbf{Z}\mathbf{V}^{(\rho)} \equiv \mathbf{\Phi}^{(\rho)}$ where matrix $\mathbf{V}^{(\rho)}$ has L^ρ columns.
3. Find $(\widehat{m}, \widehat{\rho}, \widehat{\mathbf{V}}^{(\widehat{\rho})}, \widehat{\mathbf{Z}})$ as a solution of the optimization problem (44).
4. Choose $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{Z}}\widehat{\mathbf{V}}^{(\widehat{\rho})}\mathbf{H}$ and obtain $\widehat{\mathbf{\Lambda}}$ by packing $\widehat{\mathbf{\Theta}}$ into a tensor.

Note that construction of the estimator $\widehat{\mathbf{\Lambda}}$ does not require knowledge of ν_1, ν_2, K_1 and K_2, so the estimator is fully adaptive. The following statement provides a minimax upper bound for the risk of $\widehat{\mathbf{\Lambda}}$.

THEOREM 6. *Let $\Sigma \equiv \Sigma(\nu_1, \nu_2, K_1, K_2)$ be the class of graphons satisfying Assumptions* (1), (42) *and* (43). *If $\widehat{\mathbf{\Lambda}}$ is obtained as a solution of optimization problem* (44) *as described above, then*

$$
\begin{aligned}
\sup_{f\in\Sigma} \frac{\mathbb{E}\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}^*\|^2}{n^2 L} \le C \min_{\substack{1\le h\le n-r \\ 0\le\rho\le 1}} \Bigg\{ &\frac{L^{\rho-1}}{h^{2\nu_1}} + \frac{I(\rho < 1)}{L^{2\rho\nu_2+1}} \\
&+ \frac{(h+r)^2(1 + (1-\rho)\log L)}{n^2 L^{1-\rho}} + \frac{\log(h+r)}{nL}\Bigg\},
\end{aligned} \tag{45}
$$

where the constant C in (45) *depends on ν_1, ν_2, K_1 and K_2 only.*

Note that h in (45) stands for $h = m - r$ where m is the number of blocks in the DSBM which approximates the graphon, hence, $h \le n - r$. On the other hand, $h \ge 0$ since one needs at least r blocks to approximate the graphon that

satisfies condition (42). Since the expression in the right- hand side of (45) is rather complex and is hard to analyze, we shall consider only two regimes: (a) $r = r_{n,L} \geq 2$ may depend on n and L and $\nu_1 = \infty$; or (b) $r = r_0 \geq 1$ is a fixed quantity independent of n and L. The first regime corresponds to a piecewise constant (in x and y) graphon that generates the DSBM while the second regime deals with the situation where f is a piecewise smooth function of all three arguments with a finite number of jumps. In the first case, we set $h = 2$, in the second case, we choose h to be a function of n and L. By minimizing the right-hand side of (45), we obtain the following statement.

COROLLARY 3. *Let $\widehat{\mathbf{\Lambda}}$ be obtained as a solution of optimization problem (44) as described above. Then, for $\Sigma \equiv \Sigma(\nu_1, \nu_2, K_1, K_2)$ and C independent of n and L, one has*

$$\sup_{f\in\Sigma} \frac{\mathbb{E}\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}^*\|^2}{n^2 L}$$

$$\leq \begin{cases} C \min\left\{ \frac{1}{L}\left[\left(\frac{r}{n}\right)^2 \log\left(\frac{n}{r}\right)\right]^{\frac{2\nu_2}{2\nu_2+1}}; \left(\frac{r}{n}\right)^2 \right\} + \frac{C\log r}{nL}, & r = r_{n,L}; \\ C \min\left\{ \frac{1}{L}\left(\frac{\log L}{n^2}\right)^{\frac{2\nu_1\nu_2}{(\nu_1+1)(2\nu_2+1)}}; \left(\frac{\log L}{n^2}\right)^{\frac{\nu_1}{\nu_1+1}} \right\} + \frac{C\log n}{nL}, & r = r_0. \end{cases} \tag{46}$$

In order to assess optimality of the penalized least squares estimator obtained above, we derive lower bounds for the minimax risk over the set $\Sigma(\nu_1, \nu_2, K_1, K_2)$. These lower bounds are constructed separately for each of the two regimes.

THEOREM 7. *Let matrix $\mathbf{H}$ satisfy assumptions (33) and $\nu_2 \geq 1/2$ in (43). Then, for C independent of n and L, one has*

$$\inf_{\widehat{\mathbf{\Lambda}}} \sup_{f\in\Sigma(\nu_1,\nu_2,K_1,K_2)} \mathbb{P}_{\mathbf{\Lambda}}\left\{ \frac{\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|^2}{n^2 L} \geq \Delta(n, L) \right\} \geq \frac{1}{4}, \tag{47}$$

where

$$\Delta(n, L) = \begin{cases} C \min\left\{ \frac{1}{L}\left[\left(\frac{r}{n}\right)^2\right]^{\frac{2\nu_2}{2\nu_2+1}}; \left(\frac{r}{n}\right)^2 \right\} + \frac{C\log r}{nL}, & r = r_{n,L}; \\ C \min\left\{ \frac{1}{L}\left(\frac{1}{n^2}\right)^{\frac{2\nu_1\nu_2}{(\nu_1+1)(2\nu_2+1)}}; \left(\frac{1}{n^2}\right)^{\frac{\nu_1}{\nu_1+1}} \right\} + \frac{C\log n}{nL}, & r = r_0. \end{cases} \tag{48}$$

It is easy to see that the value of $\Delta(n, L)$ coincides with the upper bound in (46) up to a at most a logarithmic factor of n/r or L. In both cases, the first quantities

in the minimums correspond to the situation where f is smooth enough as a function of time, so that application of transform $\mathbf{H}$ improves estimation precision by reducing the number of parameters that needs to be estimated. The second quantities represent the case where one needs to keep all elements of vector $\mathbf{d}$ and hence application of the transform yields no benefits. The latter can be due to the fact that ν_2 is too small or L is too low.

The upper and the lower bounds in Theorems 6 and 7 look somewhat similar to the ones appearing in anisotropic functions estimation [see, e.g., Lepski (2015)]. Note also that although in the case of a time-independent graphon ($L = 1$), the estimation precision does not improve if $\nu_1 > 1$, this is not any longer true in the case of a dynamic graphon. Indeed, the right-hand sides in (48) become significantly smaller when ν_1, ν_2 or L grow.

REMARK 4 (The DSBM and the dynamic graphon). Observe that the definition (1) of the dynamic graphon assumes that vector $\boldsymbol{\zeta}$ is independent of t. This is due to the fact that, to the best of our knowledge, the notion of the dynamic graphon with $\boldsymbol{\zeta}$ being a function of time has not yet been developed by the probability community. For this reason, we restrict our attention to the case where we are certain that, at any time point, the graphon describes the limiting behavior of the network as $n \to \infty$. Nevertheless, we believe that when the concept of the dynamic graphon is established, our techniques will be useful for its estimation.

In the case of a piecewise constant graphon, our setting corresponds to the situation where the nodes of the network do not switch their group memberships in time, so that $n_0 = 0$ in (24). Therefore, a piecewise constant graphon ($r = r_{n,L}$, $\nu_1 = \infty$) is just a particular case of the general DSBM since the latter allows any temporal changes of nodes' memberships. However, the dynamic piecewise constant graphon formulation enables us to derive specific minimax convergence rates for estimators of $\mathbf{\Lambda}$ in terms of n, L and r. On the other hand, the piecewise smooth graphon ($r = r_0$, $\nu_1 < \infty$) is an entirely different object that is not represented by the DSBM.

8. Discussion. In the present paper, we considered estimation of connection probabilities in the context of dynamic network models. To the best of our knowledge, this is the first paper to propose a fully nonparametric model for the time-dependent networks which treats connection probabilities for each group as the functional data and allows to exploit the stability in the group memberships over time. The paper derives adaptive penalized least squares estimators of the tensor of the connection probabilities in a nonasymptotic setting and shows that the estimators are indeed minimax optimal by constructing the lower bounds for the risk. This is done via vectorization technique which is very useful for the task in the paper and can be very beneficial for solution of other problems such as, for example, inference in bi-clustering models mentioned in Remark 2. In addition, we show that the correct penalty consists of two parts: the portion which accounts

for the complexity of estimation and the portion which accounts for the complexity of clustering and is proportional to the logarithm of the cardinality of the set of clustering matrices. The latter is a novel result and it is obtained by using the innovative Packing lemma (Lemma 4 in the Supplementary Material) which can be viewed as a version of the Varshamov–Gilbert lemma for clustering matrices. Finally, the methodologies of the paper allow a variety of extensions.

1. (*Inhomogeneous or nonsmooth connection probabilities*.) Assumption (43) essentially implies that probabilities of connections are spatially homogeneous and are represented by smooth functions of time that belong to the same Sobolev class. The model, however, can be easily generalized. First, by letting $\mathbf{H}$ be a wavelet transform and assuming that for any fixed x and y, function $f(x, y, \cdot)$ belongs to a Besov ball, one can accommodate the case where $f(x, y, \cdot)$ has jump discontinuities. Furthermore, by using a weaker version of condition (43), similar to how this was done in Klopp and Pensky (2015), one can treat the case where functions $f(x, y, t)$ are spatially inhomogeneous.

2. (*Time-dependent number of nodes*.) One can apply the theory above even when the number of nodes in the network changes from one time instant to another. Indeed, in this case we can form a set that includes all nodes that have ever been in the network and denote their number by n. Consider a class Ω_0 such that all nodes in this class have zero probability of interaction with each other or any other node in the network. At each time instant, place all nodes that are not in the network into the class Ω_0. After that, one just needs to modify the optimization procedures by placing additional restriction that the out-of-the-network nodes indeed belong to class Ω_0 and that $\mathbf{G}_{0,k,l} = 0$ for any $k = 0, 1, 2, \ldots, m$ and $l = 1, \ldots, L$.

3. (*Adaptivity to clustering complexity*.) Although, in the case of the DSBM, our estimator is adaptive to the unknown number of classes, it requires knowledge about the complexity of the set of clustering matrices. For example, if at most n_0 nodes can change their memberships between two consecutive time points and n_0 is a fixed quantity independent of n and m, we can replace n_0 by $\log n$ that dominates n_0 if n is large enough. However, if n_0 depends on n and m, development of an adaptive estimator would require an additional investigation.

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/18-AOS1751SUPP; .pdf). The supplement contains proofs of all statements in the paper.

REFERENCES

AMINI, A. A. and LEVINA, E. (2018). On semidefinite relaxations for the block model. *Ann. Statist.* **46** 149–179. MR3766949

ANAGNOSTOPOULOS, A., LACKI, J., LATTANZI, S., LEONARDI, S. and MAHDIAN, M. (2016). Community detection on evolving graphs. In *NIPS* 3522–3530.

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

BORGS, C., CHAYES, J. T., COHN, H. and GANGULY, S. (2016). Consistent nonparametric estimation of heavy-tailed sparse graphs. Available at arXiv:1508.06675v2.

DURANTE, D. and DUNSON, D. B. (2016). Locally adaptive dynamic networks. *Ann. Appl. Stat.* **10** 2203–2232. MR3592054

DURANTE, D., DUNSON, D. B. and VOGELSTEIN, J. T. (2017). Nonparametric Bayes modeling of populations of networks. *J. Amer. Statist. Assoc.* **112** 1516–1530. MR3750873

GAO, C., LU, Y. and ZHOU, H. H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43** 2624–2652. MR3405606

GAO, C., LU, Y., MA, Z. and ZHOU, H. H. (2016). Optimal estimation and completion of matrices with biclustering structures. *J. Mach. Learn. Res.* **17** Paper No. 161, 29. MR3569248

GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.* **18** Paper No. 60, 45. MR3687603

GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.

GUPTA, A. K. and NAGAR, D. K. (2000). *Matrix Variate Distributions*. *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics* **104**. Chapman & Hall/CRC, Boca Raton, FL. MR1738933

HAN, Q., XU, K. S. and AIROLDI, E. M. (2015). Consistent estimation of dynamic and multi-layer block models. In *Proceedings of the 32nd International Conference on Machine Learning, Lille, France* 1511–1520.

KLOPP, O. and PENSKY, M. (2015). Sparse high-dimensional varying coefficient model: Nonasymptotic minimax study. *Ann. Statist.* **43** 1273–1299. MR3346703

KLOPP, O., TSYBAKOV, A. B. and VERZELEN, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.* **45** 316–354. MR3611494

KOLACZYK, E. D. (2009). *Statistical Analysis of Network Data*: *Methods and Models*. *Springer Series in Statistics*. Springer, New York. MR2724362

KOLAR, M., SONG, L., AHMED, A. and XING, E. P. (2010). Estimating time-varying networks. *Ann. Appl. Stat.* **4** 94–123. MR2758086

LEE, M., SHEN, H., HUANG, J. Z. and MARRON, J. S. (2010). Biclustering via sparse singular value decomposition. *Biometrics* **66** 1087–1095. MR2758496

LEPSKI, O. (2015). Adaptive estimation over anisotropic functional classes via oracle approach. *Ann. Statist.* **43** 1178–1242. MR3346701

LOVÁSZ, L. (2012). *Large Networks and Graph Limits*. *American Mathematical Society Colloquium Publications* **60**. Amer. Math. Soc., Providence, RI. MR3012035

LOVÁSZ, L. and SZEGEDY, B. (2006). Limits of dense graph sequences. *J. Combin. Theory Ser. B* **96** 933–957. MR2274085

MATIAS, C. and MIELE, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1119–1141. MR3689311

MINHAS, S., HOFF, P. D. and WARDA, M. D. (2015). Relax, tensors are here: Dependencies in international processes. Available at arXiv:1504.08218.

OLHEDE, S. C. and WOLFE, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proc. Natl. Acad. Sci. USA* **111** 14722–14727.

PENSKY, M. (2019). Supplement to "Dynamic network models and graphon estimation." DOI:10.1214/18-AOS1751SUPP.

XING, E. P., FU, W. and SONG, L. (2010). A state-space mixed membership blockmodel for dynamic network tomography. *Ann. Appl. Stat.* **4** 535–566. MR2758639

XU, K. S. (2015). Stochastic block transition models for dynamic networks. In *Proceedings of the* 18*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*) 2015, *San Diego*, *CA*, *USA*, *JMLR*:*W&CP* **38**.

XU, K. S. and HERO, A. O. III (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE J. Sel. Top. Signal Process.* **8** 552–562.

YANG, T., CHI, Y., ZHU, S., GONG, Y. and JIN, R. (2011). Detecting communities and their evolutions in dynamic social networks—a Bayesian approach. *Mach. Learn.* **82** 157–189. MR3108191

ZHANG, A. Y. and ZHOU, H. H. (2016). Minimax rates of community detection in stochastic block models. *Ann. Statist.* **44** 2252–2280. MR3546450

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CENTRAL FLORIDA
ORLANDO, FLORIDA 32816-1364
USA
E-MAIL: Marianna.Pensky@ucf.edu