Blind identification of stochastic block models from dynamical observations*

Michael T. Schaub†, Santiago Segarra‡, and John N. Tsitsiklis§

Abstract. We consider a blind identification problem in which we aim to recover a statistical model of a network without knowledge of the network's edges, but based solely on nodal observations of a certain process. More concretely, we focus on observations that consist of snapshots of a diffusive process that evolves over the unknown network. We model the network as generated from an independent draw from a latent stochastic block model (SBM), and our goal is to infer both the partition of the nodes into blocks, as well as the parameters of this SBM. We present simple spectral algorithms that provably solve the partition recovery and parameter estimation problems with high accuracy. Our analysis relies on recent results in random matrix theory and covariance estimation, and associated concentration inequalities. We illustrate our results with several numerical experiments.

Key words. System identification, network inference, topology inference, consensus dynamics, stochastic block model, blind identification

AMS subject classifications. 68R10, 62M15, 60B20, 15A29, 15A18

1. Introduction.

We consider the problem of inferring a statistical model of a network with unknown adjacency matrix $\boldsymbol{A} \in \{0,1\}^{n\times n}$, based on several snapshots of a dynamical process that evolves over the network. Specifically, we are interested in the following setup (a more detailed description will be provided in Section 3): We observe s independent samples $\boldsymbol{x}_T^{(1)}, \dots, \boldsymbol{x}_T^{(s)}$ of a dynamical process at a fixed time T. Each sample is generated by the discrete-time system

$$\boldsymbol{x}_{t+1}^{(i)} = f(\boldsymbol{A})\boldsymbol{x}_t^{(i)}, \qquad i = 1, \dots, s, \tag{1.1}$$

initialized at some unknown $\boldsymbol{x}_0^{(i)}$, modeled as a random vector with unit covariance matrix ($\mathbb{E}[\boldsymbol{x}_0^{(i)}[\boldsymbol{x}_0^{(i)}]^\top] = \boldsymbol{I}$) drawn from some sub-Gaussian distribution (e.g., a standard multivariate Gaussian), and independently for each i. Furthermore, we assume that the function f describing the dynamical network process is known. Finally, $\boldsymbol{A}$ is assumed to be drawn independently according to a stochastic block model (SBM), so that $\boldsymbol{A}$ is the same, albeit unknown, for each sample $\boldsymbol{x}_T^{(i)}$. Based on such data, our task is to identify the generative model underpinning the network: we wish to infer the parameters and the latent node partition of the SBM, without observing the network edges, by relying solely on the observations $\{\boldsymbol{x}_T^{(i)}\}_{i=1}^s$ at the nodes.

1.1. Motivation.

In many applications we are confronted with the following scenario: we observe snapshots of the state of a system at particular times, and based on these observations

*Submitted to the editors DATE.

Funding: This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska- Curie grant agreement No 702410 (M. Schaub)

†Institute for Data, Systems and Society, Massachusetts Institute of Technology, USA and Department of Engineering Science, University of Oxford, UK (mschaub@mit.edu).

‡Department of Electrical and Computer Engineering, Rice University, USA (segarra@rice.edu).

§Laboratory of Information and Decision Systems, Massachusetts Institute of Technology, USA (jnt@mit.edu)

we wish to infer the nature of the (dynamical) interactions between the entities that we observe. Examples abound and can be found across different fields [17, 18, 37].

Consider, e.g., measuring some node activities such as the expression of opinions at s different, sufficiently separated instances of time in a social network. While the underlying social connections are typically not observable, the latent (unobserved) network will be essentially static, and may be assumed to have emerged according to some homophilic interactions describable by an SBM. The measured opinion profiles will be shaped by both some (random) initial personal opinions of the individuals as well as certain network effects. Similar examples may arise when measuring (structurally stable) biological systems responding to different stimuli. For instance, we might have access to a set of measurements from certain regions of the brain, and would like to infer the coupling between these regions.

The above are canonical examples of problems where the system of interest is composed of a large number of different entities and can be conveniently conceptualized as a network. In this abstraction, nodes represent the individual entities and the couplings between these nodes take place along the edges of the network. In many settings, we may have reason to believe that an observed dynamical process emerges from the interaction between nodes over such a network. While we may have some knowledge about the dynamics of each individual agent, the edges of the network – which enable the interactions between nodes – are often unknown or only partially known. In such cases it is often desirable to learn the structure of the network based on observational data. This is a general class of network inference problems, which is a fundamental constituent of network analysis, but has received limited attention.

On the other hand, there are many situations where inferring exactly the couplings in the network e.g., identifying all edges, is out of reach, for a number of reasons. First, we may not have access to sufficiently many samples. Second, the data and the network itself may be subject to some fluctuations with associated uncertainties. Finally, we may only be able to partially observe the dynamical system, so that the full model is non-identifiable.

In such cases, it is then reasonable to aim at an *approximate* description of the system at hand, in terms of a statistical model that captures the coarse-grained features of the underlying network. This is precisely the type of problem addressed in this paper, for the special case where the dynamics are of a linear, diffusive type, and the network model to be inferred is an SBM.

1.2. Contributions. We present a rigorous study of blind model identification for random networks. Our main contributions are as follows:

First, we present general algorithmic ideas on how to solve a blind identification problem using spectral methods, and introduce simple algorithms to solve the concrete problems considered in this paper (see Sections 4 and 5, and Algorithms 4.1 and 5.1).

Second, we establish a concentration inequality about the covariance matrix of the observed samples (see Theorem 4.1), which ultimately enables us to quantify the estimation error.

Finally, we discuss how this concentration result translates into bounds on the estimation errors for the parameters of the SBM, and provide explicit bounds for the recovery error of the latent partitions (see Theorems 4.2 and 5.2).

1.3. Related work. Many different network inference notions have featured in the literature, ranging from estimating "functional" couplings, based on statistical association measures such

as correlation or mutual information [12,25], all the way to causal inference [36]. The notion of inference that we consider here is what may be called "topological" inference: given a system of dynamical units, we want to infer their direct "physical" interactions [37]. More specifically, for an evolving dynamical process we would like to infer the underlying adjacency matrix. This may be viewed as a system identification problem, as considered in Control Theory [17,18]. However, in the system identification literature, one can usually control the inputs to the system and observe the full system response. This is different from our setting, where there is no control, and we only observe a limited number of snapshots.

Problems of network identification that are somewhat similar to ours have received wide interest in the literature recently, using techniques from optimization, spectral analysis, and statistics [4,8,9,15,21–23,30,32,33,35,40,45]. In particular, the case of diffusive dynamics, similar to ours, is considered in [34] and [35], which aim at inferring the network topology based on the (cross-) power spectral density of a consensus process driven by noise, and a node knockout strategy. Methods combining spectral identification with optimization techniques are studied in [9,10,32,45].

In all of the above works, however, the focus is on inferring the *exact* network topology, i.e., the complete specification of the network in terms of its adjacency or Laplacian matrices. In contrast, our target is not to infer a specific network coupling, but rather to identify a statistical model thereof. Stated differently, our goal is to infer a distribution over networks that is compatible with the available observations. This scenario is particularly relevant if our ability to reconstruct the network is limited by the amount of samples that we can obtain. Closest to our approach here is the conference paper [29], in which a related observational model is discussed with a new adjacency matrix drawn for each sample in (1.1), and which provides results about the asymptotic consistency of partition recovery. In contrast, we provide non-asymptotic results in the number of samples and, in addition, we discuss the problem of parameter estimation. The works [41,42] focus on recovering the network partition for low-rank excitation signals, but do not aim to infer a generative model. The work by Hoffmann et al. [11] discusses related ideas; it presents a factor model and a Bayesian inference scheme, but does not provide any theoretical analysis.

Finally, the mathematical machinery that we use has connections to ideas from spectral methods for clustering on graphs and SBM recovery [26,27,39], and is related to the statistics literature on so-called "spiked" random matrices, with a low-rank signal component. In particular, our development relies on recent results regarding the efficiency of estimating low-rank covariance matrices [3].

1.4. Outline. The rest of the paper is structured as follows. After briefly reviewing relevant preliminaries and notation in Section 2, in Section 3 we formally introduce the identification problem that we will study. We then present our main results and algorithmic ideas on recovering the latent *partitions* of the SBM in Section 4. Building on those insights, in Section 5 we discuss procedures for learning the *parameters* of the SBM. Section 6 contains the proofs of our main technical results, and in Section 7 we present computational experiments to numerically assess the performance of our methods. We conclude with a brief discussion, and outline possible avenues for future work.

2. Preliminaries.

2.1. Networks. An undirected network $\mathcal{G}$ consists of a node set $\mathcal{V}$ of known cardinality n, an edge set $\mathcal{E}$ of unordered pairs of elements of $\mathcal{V}$, and edge weights $A_{ij} \in \{0,1\}^{n\times n}$ such that $A_{ij} = A_{ji} = 1$ if and only if $(i,j) \in \mathcal{E}$. The edge weights A_{ij} can be conveniently collected as entries of a symmetric adjacency matrix $\boldsymbol{A}$. We define the diagonal degree matrix $\boldsymbol{D} = \operatorname{diag}(\boldsymbol{A}\mathbf{1})$, and then define the normalized adjacency matrix as $\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$. Note that $\boldsymbol{L}$ is a shifted version of the normalized Laplacian matrix of the graph, and is sometimes referred to in the literature simply as the "normalized Laplacian" [26, 27].

2.2. The stochastic block model. The SBM is a latent variable model that defines a probability measure over the set of undirected networks of fixed size n, represented by an adjacency matrix $\boldsymbol{A} \in \{0,1\}^{n\times n}$. In an SBM, the network is assumed to be partitioned into k groups of nodes, so that each node i is endowed with a latent group label $g_i \in \{1,\dots,k\}$. Given these latent group labels, each link A_{ij} is an independent Bernoulli random variable that takes value 1 with probability Ω_{g_i,g_j} and value 0 otherwise; that is, the probability of a link between two nodes depends only on their group labels:

$$A_{ij}|g_i,g_j \sim \operatorname{Ber}(\Omega_{g_i,g_j}).$$

To compactly describe the model, we collect link probabilities between the different groups in a symmetric affinity matrix $\boldsymbol{\Omega} = [\Omega_{ij}] \in [0,1]^{k\times k}$. Furthermore we define a partition indicator matrix $\boldsymbol{G} \in \{0,1\}^{n\times k}$, with entries $G_{ij} = 1$ if node i is in class j and $G_{ij} = 0$ otherwise. Based on these definitions, we can write the expected adjacency matrix under the SBM, given the partition and the class labels as[1]

$$\boldsymbol{\mathcal{A}} := \mathbb{E}[\boldsymbol{A} \,|\, \boldsymbol{G}] = \boldsymbol{G}\boldsymbol{\Omega}\boldsymbol{G}^\top. \tag{2.1}$$

As $\boldsymbol{\mathcal{A}}$ encodes all the parameters of an SBM, we will in the sequel refer to such a stochastic block model as $\mathcal{M}(\boldsymbol{\mathcal{A}})$.

Equation (2.1) shows that, in expectation, the SBM leads to a network that has an adjacency matrix with a low-rank block structure. Due to this special structure, there exists a simple relationship between the partition indicator matrix $\boldsymbol{G}$ and the eigenvectors of the normalized version of the expected adjacency matrix

$$\boldsymbol{\mathcal{L}} := \mathbb{E}[\boldsymbol{D}]^{-1/2}\mathbb{E}[\boldsymbol{A}]\mathbb{E}[\boldsymbol{D}]^{-1/2},$$

which is given in Lemma 2.1 below, and proved in the Appendix. We denote the eigenvalues of $\boldsymbol{\mathcal{L}}$ by μ_i, and assume that they are sorted in order of decreasing magnitude, so that $|\mu_1| \geq |\mu_2| \geq \cdots \geq |\mu_k| > |\mu_{k+1}| = \cdots = |\mu_n| = 0$.

Lemma 2.1. *Consider the matrix $\boldsymbol{V} = [\boldsymbol{v}_1,\dots,\boldsymbol{v}_k]$ of eigenvectors of $\boldsymbol{\mathcal{L}}$ corresponding to the k nonzero eigenvalues. Then, there exists an orthogonal matrix $\boldsymbol{\mathcal{U}}$ such that*

$$\boldsymbol{V} = \boldsymbol{G}(\boldsymbol{G}^\top\boldsymbol{G})^{-1/2}\boldsymbol{\mathcal{U}}. \tag{2.2}$$

[1] Note the slight abuse of notation here with respect to the conditional expectation. We treat the partition indicator matrix $\boldsymbol{G}$ as a degenerate random variable with a deterministic, but unknown value.

Moreover, if we define the row-normalized eigenvector matrix $\widetilde{\boldsymbol{V}} = \mathrm{diag}(\boldsymbol{V}\boldsymbol{V}^\top)^{-1/2}\boldsymbol{V}$*, then*

$$\widetilde{\boldsymbol{V}} = \boldsymbol{G}\boldsymbol{\mathcal{U}}. \tag{2.3}$$

Given an SBM with an expected adjacency matrix $\boldsymbol{\mathcal{A}}$ as described above, let us denote the expected degree of node i by $\delta_i := \mathbb{E}[\boldsymbol{D}_{ii}] = [\boldsymbol{\mathcal{A}}\mathbf{1}]_i$, and let $\delta_{\min}$ be the smallest expected degree. We define the following class of SBMs with n nodes.

Definition 2.2 (SBMs with concentration property). *Consider an SBM* $\mathcal{M}(\boldsymbol{\mathcal{A}})$ *with* n *nodes and minimal expected degree* δ_{min}*. We define the following set of SBMs:*

$$\mathcal{M}_n(\epsilon) := \left\{ \mathcal{M}\left(\boldsymbol{\mathcal{A}}\right) \mid \delta_{min} > 3\left(1 + \ln\left(\frac{4}{\epsilon}\right)\right)\ln(n) \right\}$$

The condition on the minimal degree essentially restricts $\mathcal{M}_n(\epsilon)$ to correspond to models that generate connected graphs with high probability. Indeed, if we were to consider an Erdős-Rényi graph, it is well known that if $\delta_{\min} > c\ln(n)$, with $c > 1$, a random graph is connected with high probability, as $n \to \infty$. If we were to observe the graph sampled from an SBM, the exact recovery of the partition with high-probability is indeed only possible under this type of condition [1].

For the above class of models the following concentration result can be established.

Lemma 2.3 (Concentration for normalized adjacency matrix). *Fix an arbitrary* $\epsilon > 0$*. Consider a normalized adjacency matrix* $\boldsymbol{L} := \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$ *computed from an adjacency matrix* $\boldsymbol{A}$ *generated according to an SBM model* $\mathcal{M}(\boldsymbol{\mathcal{A}})$*. Then, if* $\mathcal{M}(\boldsymbol{\mathcal{A}}) \in \mathcal{M}_n(\epsilon)$ *the following bound holds with probability at least* $1 - \epsilon$*:*

$$\|\boldsymbol{L} - \boldsymbol{\mathcal{L}}\| \leq 3\sqrt{\frac{3\log(4n/\epsilon)}{\delta_{min}}},$$

where we recall that $\boldsymbol{\mathcal{L}} := \mathbb{E}[\boldsymbol{D}]^{-1/2}\mathbb{E}[\boldsymbol{A}]\mathbb{E}[\boldsymbol{D}]^{-1/2}$*.*

Lemma 2.3 follows from a result by Chung and Radcliffe [5]. The proof, which parallels [5], is given in the Appendix for completeness.

2.3. Sub-Gaussian random vectors. For convenience, we recall here the definition of sub-Gaussian random variables and vectors.

Definition 2.4 (Sub-Gaussian random variables and vectors). *A zero-mean random variable* $x \in \mathbb{R}$ *is* sub-Gaussian *if there exists a constant* $\sigma > 0$ *such that* $\mathbb{E}[\exp(tx)] \leq \exp(t^2\sigma^2/2)$*, for all* $t \in \mathbb{R}$*. A zero-mean random vector* $\boldsymbol{x} \in \mathbb{R}^n$ *is sub-Gaussian if for any non-random* $\boldsymbol{u} \in \mathbb{R}^n$*, the random variable* $z = \boldsymbol{u}^\top\boldsymbol{x}$ *is sub-Gaussian.*

Note that any Gaussian random variable or vector is sub-Gaussian. We also define the sub-Gaussian norm as follows.

Definition 2.5 (Sub-Gaussian norm [3]). *The sub-Gaussian norm of a sub-Gaussian random variable* x *is* $\|x\|_{\Psi_2} := \sup_{k\geq 1} k^{-1/2}(\mathbb{E}|x|^k)^{1/k}$*. For a sub-Gaussian random vector* $\boldsymbol{x}$*, the sub-Gaussian norm* $\|\boldsymbol{x}\|_{\Psi_2}$ *is defined as* $\|\boldsymbol{x}\|_{\Psi_2} := \sup_{\boldsymbol{u}\in\mathbb{R}^n\setminus\{0\}} \|\boldsymbol{u}^\top\boldsymbol{x}\|_{\Psi_2}/\|\boldsymbol{u}\|$*.*

3. The blind identification Problem. We now introduce the specific inference problem that we will study. We consider the dynamical model described in (1.1), for a specific choice of the mapping f, although most of our algorithmic ideas can be extended to more general functions f, e.g., those describing a low-pass graph filter, as considered in [29, 42]. The details are as follows.

1. We have a discrete-time dynamical system with state vector $\boldsymbol{x}_t \in \mathbb{R}^n$, which evolves according to

$$\boldsymbol{x}_{t+1} = \boldsymbol{L}\boldsymbol{x}_t, \tag{3.1}$$

   where $\boldsymbol{L}$ is the (unknown) normalized adjacency matrix of the (unobserved) underlying graph. This corresponds to (1.1) for the specific case $f : \boldsymbol{A} \mapsto \boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$.
2. We assume that the normalized adjacency matrix $\boldsymbol{L}$ is generated by first drawing a graph from an SBM with k blocks, and some (unknown) affinity matrix $\boldsymbol{\Omega}$, and then normalizing its adjacency matrix $\boldsymbol{A}$ to obtain $\boldsymbol{L}$.
3. We obtain s samples $\boldsymbol{x}_T^{(i)}$, $i = 1, \dots, s$, at a known time T.
4. The (unobserved) initial vectors $\boldsymbol{x}_0^{(i)}$ for the different samples are i.i.d., drawn from a zero-mean sub-Gaussian vector distribution, where the components of the vector are uncorrelated, i.e., $\mathbb{E}[\boldsymbol{x}_0^{(i)}[\boldsymbol{x}_0^{(i)}]^\top] = \boldsymbol{I}$.

The discrete-time dynamics in (3.1) are closely related to diffusion processes on networks [20, 31]. Indeed, $\boldsymbol{L}$ is related via a similarity transformation to the transition matrix $\boldsymbol{A}\boldsymbol{D}^{-1}$ of an unbiased random walk on the network. Such random walks have been used to model a variety of processes, such as the spreading of rumors and opinions, and have found applications in other areas of network analysis such as ranking nodes or detecting communities [20]. Another interpretation of the model is in terms of (discrete-time) consensus processes, which have been considered extensively in the Control Theory literature [14, 24, 38].

Note that if the number of available samples becomes large ($s \to \infty$), we could in principle try to infer the exact adjacency matrix $\boldsymbol{A}$. However, we are interested in a situation where s is comparatively small ($s \ll n$), and our goal is to infer the SBM model that generated $\boldsymbol{A}$ (and thus $\boldsymbol{L}$), that is, the latent partition and the affinity matrix $\boldsymbol{\Omega}$.

We specifically consider the following two problems, which we refer to as the partition recovery and the parameter estimation problems.

Problem 3.1. *(SBM partition recovery from dynamics). Given s independent samples, each one drawn at a fixed time T according to the above described model, recover the latent partition of the nodes into blocks.*

Problem 3.2. *(SBM parameter estimation from dynamics). Given s independent samples, each one drawn at a fixed time T according to the above described model, recover the affinity matrix $\boldsymbol{\Omega}$ describing the link probabilities in the underlying SBM.*

Note that, as the assignment of labels to groups is arbitrary, we can only hope to recover the block structure and the affinity matrix $\boldsymbol{\Omega}$ up to a permutation of the group labels. Moreover, it should be apparent that the difficulty of the above problems will depend on the specific realization of the initial conditions $\boldsymbol{x}_0^{(i)}$ and of the adjacency matrix $\boldsymbol{A}$. Thus, we cannot expect accurate inference for every possible realization. Instead we look for algorithms and results

Algorithm 4.1 Partition recovery

Input: Samples $\{\boldsymbol{x}_T^{(i)}\}_{i=1}^s$, Number of groups k
Output: Partition $\mathcal{P}$

1: Compute sample covariance $\widehat{\boldsymbol{\Sigma}}(T) := \frac{1}{s}\sum_{i=1}^s [\boldsymbol{x}_T^{(i)} - \bar{\boldsymbol{x}}_T][\boldsymbol{x}_T^{(i)} - \bar{\boldsymbol{x}}_T]^\top$
2: Compute top k eigenpairs $|\lambda_1| \geq |\lambda_2| \geq \ldots : \{(\lambda_i, \boldsymbol{w}_i)\}_{i=1}^k \leftarrow \mathrm{eig}(\widehat{\boldsymbol{\Sigma}}(T))$
3: Form the matrix $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k] \in \mathbb{R}^{n\times k}$ and normalize its rows to have unit norm $\widetilde{\boldsymbol{W}} = \mathrm{diag}(\boldsymbol{W}\boldsymbol{W}^\top)^{-1/2}\boldsymbol{W}$
4: Perform k-means clustering on the rows of $\widetilde{\boldsymbol{W}}$: $\mathcal{P} \leftarrow k\text{-means}(\widetilde{\boldsymbol{W}}^\top)$

that, under certain assumptions, guarantee accurate inference with high probability.

We make the following technical assumption about the data generation process.

Assumption 3.3 (Bounded moments assumption). *We assume that there exists an absolute constant $c_0 > 0$ such that $\mathbb{E}[(\boldsymbol{u}^\top \boldsymbol{x}_0^{(i)})^2] \geq c_0 \|\boldsymbol{u}^\top \boldsymbol{x}_0^{(i)}\|_{\Psi_2}^2$, for all $\boldsymbol{u}$.*

As discussed in [3], Assumption 3.3 effectively bounds the moments of $\boldsymbol{x}_0^{(i)}$ as a function of the second moment of $\boldsymbol{x}_0^{(i)}$. As a special case, this assumption is satisfied by a standard Gaussian vector, with $c_0 = \pi/2$ [3]. Assumption 3.3 will enable us to invoke (via Lemma 6.2) the results of [3] on the estimation of low-rank covariance matrices, thereby facilitating our analysis.

4. Partition recovery.

In this section we first present our main results and an algorithm for solving the partition recovery problem (Problem 1) and then provide a broader discussion of the underlying algorithmic ideas.

4.1. Main results.

Our algorithm for partition recovery relies on spectral clustering, and is displayed as Algorithm 4.1. It builds upon the idea that the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}(T)$ at time T, defined by

$$\widehat{\boldsymbol{\Sigma}}(T) := \frac{1}{s}\sum_{i=1}^s [\boldsymbol{x}_T^{(i)} - \bar{\boldsymbol{x}}_T][\boldsymbol{x}_T^{(i)} - \bar{\boldsymbol{x}}_T]^\top, \tag{4.1}$$

where $\bar{\boldsymbol{x}}_T = \frac{1}{s}\sum_{i=1}^s \boldsymbol{x}_T^{(i)}$ is the sample mean, can serve as a proxy for what we call the "population covariance matrix" $\boldsymbol{S}(T) := \boldsymbol{\mathcal{L}}^{2T}$.

Because the initial vector has identity covariance, it is seen that $\boldsymbol{S}(T)$ would be the true covariance matrix at time T if the dynamics were evolving according to the matrix $\boldsymbol{\mathcal{L}}$. Intuitively, the matrix $\boldsymbol{\mathcal{L}}$ encodes the partition structure of the SBM, and since $\boldsymbol{S}(T)$ is simply a matrix power of $\boldsymbol{\mathcal{L}}$ it has the same dominant eigenspace. Being able to compute a close proxy for $\boldsymbol{S}(T)$ enables us to gain information about the partition structure. Thus, as long as the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}(T)$ is close to the population covariance matrix $\boldsymbol{S}(T)$, we can expect good recovery performance. Theorem 4.1 below adresses the relation between these two covariance matrices.

Theorem 4.1. *Let $\boldsymbol{L}$ be the normalized adjacency matrix constructed from a graph drawn from an SBM $\mathcal{M}(\boldsymbol{\mathcal{A}}) \in \mathcal{M}_n(\epsilon)$ with n nodes and minimal degree δ_{min}. Let $\boldsymbol{x}_0^{(i)}$ $i = 1, \ldots, s$, be i.i.d. zero-mean sub-Gaussian vectors that satisfy Assumption 3.3, with covariance matrix $\boldsymbol{I}$.*

Then, with probability at least $1 - 5s^{-1} - \epsilon$*, we have*

$$\left\| \widehat{\boldsymbol{\Sigma}}(T) - \boldsymbol{S}(T) \right\| \leq 2TM + B(T), \tag{4.2}$$

where $\widehat{\boldsymbol{\Sigma}}(T)$ *is defined in* (4.1),

$$M := 3\sqrt{\frac{3\log(4n/\epsilon)}{\delta_{min}}}, \qquad B(T) := C\left(1+\sum_{i=2}^{k}(|\mu_i|+M)^{2T} + (n-k)M^{2T}\right)\sqrt{\frac{\ln s}{s}},$$

C *is some absolute constant, and where* μ_i *are the eigenvalues of* $\boldsymbol{\mathcal{L}}$ *ordered by decreasing magnitude.*

The proof of Theorem 4.1 is given in Subsection 6.1.

For some interpretation, note that the term $2TM$ in (4.2) is independent of the number of samples. As we will see below, this term arises because the graph on which the dynamics takes place is only a proxy for the expected graph under the SBM and is drawn only once. We may think of this error term as a 'baseline' insurmountable error that we will incur no matter how many samples we can obtain from our system. While for a large enough $\delta_{\min}$ the error M can be small (cf. Lemma 2.3), the error bound increases (linearly) with time. Hence, the later we make our observations (larger T), the worse our bound from this first term becomes. From the perspective of our model (3.1) this is intuitive: as $\lambda_i(\boldsymbol{L}) \leq 1$ for all i, the relevant signal that we observe will be attenuated over time, making it more difficult to extract information about the embedded partition from our observations.

In contrast to the first term, the second term $B(T)$ in (4.2) depends both on the time T and the number of samples s. The sample dependence is related to the convergence rate of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}(T)$ to the true covariance. The time-dependence of $B(T)$ is again related to M, but in a non-linear fashion. Since $|\mu_i| < 1$ for $i = 2, \dots, k$, we see that for a small $M \ll 1$ the error term $B(T)$ decreases with time.

This leads to a trade-off: for later sampling times, we require fewer samples to achieve a good bound. More precisely, for $M \ll 1$, if the number of samples s and the time T are comparably small, then $2TM \ll (n-k)M^{2T}$ and thus $B(T)$ dominates in (4.2). However, the exponential dependence on T will reduce the terms in $B(T)$ very fast, and accordingly $2TM$ will become the dominant factor in the bound for larger T. This error reduction scales much faster than that obtained by collecting more samples (larger s).

Based on the concentration result in Theorem 4.1, we can establish the following guarantee for the partition recovery problem. This result is phrased in terms of the misclassification rate $q := |\mathbb{M}|/n$, where $\mathbb{M}$ is the set of misclassified nodes (see Subsection 6.2 for a formal definition).

Theorem 4.2. *Let* $\xi_1 \geq \cdots \geq \xi_k$ *be the* k *positive eigenvalues of the population covariance matrix* $\boldsymbol{S}(T) = \boldsymbol{\mathcal{L}}^{2T}$*. Let* $\boldsymbol{W}$ *and* $\boldsymbol{V}$ *denote the matrices formed by the top* k *eigenvectors of* $\widehat{\boldsymbol{\Sigma}}(T)$ *and* $\boldsymbol{S}(T)$*, respectively, and denote their row-normalized versions by* $\widetilde{\boldsymbol{W}}$ *and* $\widetilde{\boldsymbol{V}}$*. Define* τ *to be the minimum 2-norm of any row of* $\boldsymbol{W}$ *and* $\boldsymbol{V}$*. Then, with probability* $1 - 5s^{-1} - \epsilon$ *the following bound for the misclassification rate* q *holds:*

$$q \leq \frac{64k}{n\tau^2\xi_k^2}(2TM + B(T))^2, \tag{4.3}$$

where $B(T)$ is as defined in Theorem 4.1.

Theorems 4.1 and 4.2 provide performance guarantees that show the interplay between the number of groups k, the network size n, and the spectral properties of the network, as encapsulated in μ_k and $\|\widehat{\boldsymbol{\Sigma}}(T) - \boldsymbol{S}(T)\|$. The proof is given in Subsection 6.2.2.

Before continuing, let us make a few more remarks on some noteworthy aspects of these results.

1. The time T at which the samples are obtained is not a required input to the algorithm (cf. Algorithm 4.1). However, as is apparent from Theorems 4.1 and 4.2, the partition recovery performance is influenced by the sampling time.
2. In Algorithm 4.1, we require the number of blocks k to be an input to our algorithm. In principle, we can try to infer this parameter from the data, e.g., by looking at some appropriate statistics of the eigenvalue gaps; see for example [28, 39, 43], or the discussion in [3] on the detection of eigenvalue jumps in covariance matrix estimation. However, we will not carry out this line of analysis in this paper.
3. The same approach can be used with other types of (linear) operators derived from a low-rank statistical network model such as the SBM. For instance one may consider inference for other latent graph models, such as the degree-corrected SBM [6, 16].

4.2. Algorithm and proof ideas. We now provide a more detailed discussion of the proposed algorithm for partition recovery from a high-level perspective. The proofs in Section 6 will make these ideas mathematically precise.

The intuition underpinning our algorithm is that while we cannot observe the edges of the network, the network structure affects the second moments of the observed samples in ways that we can exploit.

To this end, we first need to establish that the empirical covariance $\widehat{\boldsymbol{\Sigma}}(T)$ serves as a good estimator of these moments. By leveraging that the effective rank $r_e(\boldsymbol{\Sigma}(T)) = \operatorname{trace}(\boldsymbol{\Sigma}(T))/\|\boldsymbol{\Sigma}(T)\|$ of the covariance matrix $\boldsymbol{\Sigma}(T)$ is low, it can be shown that $\widehat{\boldsymbol{\Sigma}}(T)$ indeed concentrates around $\boldsymbol{\Sigma}(T)$ even for relatively small number of samples (Lemma 6.1).

Second, since we assume the initial conditions $\boldsymbol{x}_0^{(i)}$ to be zero-mean sub-Gaussian with $\mathbb{E}[\boldsymbol{x}_0\boldsymbol{x}_0^\top] = \boldsymbol{I}$, the covariance matrix of our samples reflects the network structure as follows:

$$\boldsymbol{\Sigma}(T) := \mathbb{E}_{\boldsymbol{x}_0}[\boldsymbol{x}_T\boldsymbol{x}_T^\top] - \mathbb{E}_{\boldsymbol{x}_0}[\boldsymbol{x}_T]\mathbb{E}_{\boldsymbol{x}_0}[\boldsymbol{x}_T^\top] = [\boldsymbol{L}^T][\boldsymbol{L}^T]^\top = \boldsymbol{L}^{2T}, \tag{4.4}$$

where the last equality follows from the fact that the network is symmetric. Clearly, the eigenvectors of $\boldsymbol{\Sigma}(T)$ are the same as the eigenvectors of $\boldsymbol{L}$, although they correspond now to the scaled eigenvalues $\lambda_i(\boldsymbol{\Sigma}(T)) = \lambda_i^{2T}(\boldsymbol{L})$. This transformation does not change the ordering of the eigenvalues in terms of their absolute value and thus the dominant subspaces of $\boldsymbol{\Sigma}(T)$ and $\boldsymbol{L}$ are the same.

Finally, we need to show that the eigenspace of the covariance matrix $\boldsymbol{\Sigma}(T)$ contains the relevant information to recover the partition structure. To this end, we need to establish that $\boldsymbol{L}$ concentrates around $\boldsymbol{\mathcal{L}}$ and thus the corresponding eigenspaces are close. Indeed, this is implied by Lemma 2.3 for the considered model class.

From the above discussion we note that Algorithm 4.1 is subject to two sources of randomness: not only is the normalized adjacency matrix $\boldsymbol{L}$ of the network an approximation of its population version $\boldsymbol{\mathcal{L}}$, but also the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}(T)$ is only an approximation

of the true covariance matrix $\mathbf{\Sigma}(T)$. Theorem 4.1 allows us to bound both of these effects, which we then leverage in Theorem 4.2 to derive performance guarantees.

5. Parameter estimation. To estimate the *parameters* of the underlying SBM we can use a similar strategy as for the recovery of the *partitions*. The population covariance matrix $\boldsymbol{S}(T) = \boldsymbol{\mathcal{L}}^{2T}$ is a transformation of the normalized adjacency matrix. Hence, it does not only contain relevant information about the partitions, but also about the model parameters. In principle, we can then use the sample covariance matrix $\widehat{\mathbf{\Sigma}}(T)$ as an approximation of $\boldsymbol{S}(T)$ and leverage the functional dependency of $\boldsymbol{S}(T)$ on the latent parameters to estimate them. In particular, we can express the eigenvalues of $\boldsymbol{S}(T)$ as a function of the model parameters, in order to obtain a set of equations $\lambda(\boldsymbol{S}(T)) = F_\lambda(\mathbf{\Omega})$ for the parameters $\mathbf{\Omega}$. Notice that these equations are independent of the group labels of the nodes. The reason is that a permutation of the nodes defines a similarity transformation of $\boldsymbol{\mathcal{L}}$, and hence the eigenvalues of $\boldsymbol{\mathcal{L}}$ – and accordingly those of $\boldsymbol{S}(T)$ – are not affected by the ordering of the rows and columns of the matrix.

To make these ideas concrete let us consider the following example.

Example 5.1. *Consider a special case of the SBM, known as the planted partition model. Conditional on the class labels, the planted partition model with n nodes is described by probabilities a/n and b/n of linking to a node in the same class or a different class, respectively.*[2] *The affinity matrix for such a model, with k blocks, is of the form $\mathbf{\Omega} = \frac{(a-b)}{n}\boldsymbol{I}_k + \frac{b}{n}\mathbf{1}_k\mathbf{1}_k^\top$, where $\boldsymbol{I}_k$ and $\mathbf{1}_k$ denote the k-dimensional identity matrix and the vector of all ones, respectively. This leads to an expected matrix $\boldsymbol{\mathcal{L}}$ with entries*

$$[\boldsymbol{\mathcal{L}}]_{ij} = \begin{cases} \frac{a}{n} \cdot \frac{k}{a+(k-1)b}, & \text{if } g_i = g_j, \\ \frac{b}{n} \cdot \frac{k}{a+(k-1)b}, & \text{if } g_i \neq g_j, \end{cases}$$

where, we recall, g_i represents the latent group label of node i.

Based on these calculations we can find expressions for $\boldsymbol{S}(T) = \boldsymbol{\mathcal{L}}^{2T}$ as follows

$$[\boldsymbol{\mathcal{L}}^{2T}]_{ij} = \begin{cases} \frac{k-1}{n}\left(\frac{a-b}{a+(k-1)b}\right)^{2T} + \frac{1}{n} & \text{if } g_i = g_j, \\ -\frac{1}{n}\left(\frac{a-b}{a+(k-1)b}\right)^{2T} + \frac{1}{n} & \text{if } g_i \neq g_j. \end{cases} \tag{5.1}$$

In terms of spectral analysis, the nonzero eigenvalues of $\boldsymbol{S}(T)$ are:

$$\lambda_i = \begin{cases} 1, & \text{if } i = 1, \\ \left(\frac{a-b}{a+(k-1)b}\right)^{2T}, & \text{if } i = 2, \dots, k. \end{cases} \tag{5.2}$$

Notice that the equations in (5.2) do not enable us to identify the model parameters a and b. However, this is an artifact of the symmetry properties of the normalized adjacency matrix $\boldsymbol{L}$ and not of the parameter estimation strategy. In particular, this degeneracy arises

[2] In Theorem 5.2 below, we will assume that a, b and n are such that $\mathcal{M}(\boldsymbol{\mathcal{A}}) \in \mathcal{M}_n(\epsilon)$. In the context of this example this means a and b are assumed to be of order $\log(n)$.

from an invariance in the dynamical model (3.1). Indeed, if in Example 5.1 we replace (a, b) by $(\gamma a, \gamma b)$, where γ is a positive constant, the expected matrix $\boldsymbol{\mathcal{L}}$ remains the same. Thus the model parameters are not recoverable in our example, even if we had access to the full matrix $\boldsymbol{\mathcal{L}}$. In the context of our specific model (3.1), we thus have to make some additional assumptions in order to estimate all the model parameters, as discussed next.

5.1. Algorithm for parameter estimation. Despite the difficulties identified in Example 5.1, we can still obtain a viable parameter estimation approach if we assume some additional information on the model parameters. More concretely, suppose that we know of some additional constraints on $\boldsymbol{\Omega}$ (or alternatively $\boldsymbol{\mathcal{A}}$), of the form $h(\boldsymbol{\Omega}) = 0$, for some known function h.[3] Such constraints could be, e.g., on the average density of the network.

As discussed earlier, based on our dynamical model we can obtain relations of the form $\lambda(\boldsymbol{S}(T)) = F_\lambda(\boldsymbol{\Omega})$. Taking these equations and the constraints together, we obtain a system of (possibly redundant) equations of the form

$$\lambda(\boldsymbol{S}(T)) = F_\lambda(\boldsymbol{\Omega}) \quad \text{subject to} \quad h(\boldsymbol{\Omega}) = 0. \tag{5.3}$$

Our approach is to solve this system of equations, where we replace $\boldsymbol{S}(T)$ by its estimate $\widehat{\boldsymbol{\Sigma}}(T)$. We note that this system will generally have more equations than unknowns and, in the presence of noise, it will be inconsistent. There are many ways of obtaining an approximate solution in such cases. Which approach would have better statistical properties in the face of the noise present in $\widehat{\boldsymbol{\Sigma}}(T)$ is an interesting question. We do not go into such details, but provide instead a generic "meta"-algorithm.

In the description of Algorithm 5.1, the word "solve" is to be understood as some not fully specified approach to obtain an approximate solution of the system (5.3). Finally, observe that the sampling time T enters explicitly in the functional relations that we exploit, and therefore, unlike the partition recovery problem (cf. Algorithm 4.1), T must be known.

Algorithm 5.1 Model parameter estimation

Input: Samples $\{\boldsymbol{x}_T^{(i)}\}_{i=1}^s$, Number of groups k, Time T
Output: Model Parameters $\boldsymbol{\Omega}$
1: Compute the sample covariance $\widehat{\boldsymbol{\Sigma}}(T) := \frac{1}{s}\sum_{i=1}^s [\boldsymbol{x}_T^{(i)} - \bar{\boldsymbol{x}}_T][\boldsymbol{x}_T^{(i)} - \bar{\boldsymbol{x}}_T]^\top$
2: Compute top k eigenvalues (according to magnitude): $\{\lambda_i\}_{i=1}^k \leftarrow \mathrm{eig}(\widehat{\boldsymbol{\Sigma}}(T))$
3: "Solve" the set of equations (5.3) to estimate model parameters $\boldsymbol{\Omega}$

5.2. Theoretical guarantees. We can leverage our concentration result in Theorem 4.1 to provide a bound between the estimated eigenvalues $\lambda_i(\widehat{\boldsymbol{\Sigma}}(T))$ and their population counterparts $\mu_i(\boldsymbol{S}(T))$. These bounds then translate immediately into (T-dependent) error-bounds on the parameter recovery.

We illustrate this idea here by deriving a result for the planted partition model considered in Example 5.1 subject to a density contraint on the network and the assumption that $a > b$.

[3] We remark that for other dynamical models such additional assumptions may not be necessary.

Recall from Example 5.1 that, for the second eigenvalue of $\boldsymbol{\mathcal{L}}^{2T}$, the relation

$$(\lambda_2(\boldsymbol{\mathcal{L}}^{2T}))^{1/2T} = \frac{a-b}{a+(k-1)b}$$

holds. Given $\rho = \frac{a+(k-1)b}{nk}$ we can compute the parameter a for $k=2$ as

$$a = (\lambda_2(\boldsymbol{\mathcal{L}}^{2T}))^{1/2T}\rho n + \rho n. \tag{5.4}$$

To estimate a, we can thus replace $\boldsymbol{\mathcal{L}}^{2T}$ in (5.4) by the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}(T)$. For this estimation strategy we can derive the following error bound.

Theorem 5.2. *Consider an SBM $\mathcal{M}(\boldsymbol{\mathcal{A}}) \in \mathcal{M}_n(\epsilon)$ with two equally sized blocks, and a parameter matrix of the form $\boldsymbol{\Omega} = \frac{(a-b)}{n}\boldsymbol{I}_2 + \frac{b}{n}\mathbf{1}_2\mathbf{1}_2^\top$ with $a > b$. Assume $\rho = (a+b)/2n$ is given. Define the estimator of a given by*

$$\hat{a} = (\lambda_2(\widehat{\boldsymbol{\Sigma}}(T)))^{1/2T}\rho n + \rho n. \tag{5.5}$$

Then, using the notation of Theorem 4.1, with probability $1 - 5s^{-1} - \epsilon$ the estimation error $\eta := |\hat{a} - a|/a$ is upper bounded as:

$$\eta \leq \frac{(2TM + B(T))^{1/2T}}{\lambda_2(\boldsymbol{\mathcal{L}}) + 1}. \tag{5.6}$$

The proof of Theorem 5.2 is given in Section 6.3. From the numerator in (5.6) it can be seen that the sampling time T plays a fundamental role in the estimation accuracy of $\hat{a}$. Moreover, this dependency is similar to the one encountered in Theorem 4.2 for the partition recovery. In particular, there is a term that increases linearly with T, i.e. $2TM$, and another term that may decrease with T, i.e., $B(T)$ (cf. the discussion after Theorem 4.1). The denominator in (5.6) can also be explained intuitively. A larger $\lambda_2(\boldsymbol{\mathcal{L}})$ is associated with a marked difference between a and b [cf. (5.2)], which translates into better delineated blocks within our SBM. In this scenario, estimating a should be easier, and this is captured in (5.6).

Before moving to a detailed description of our theoretical analysis in Section 6, we share the following remark about an alternative avenue to estimate the SBM parameters.

Remark 5.3 (Parameter estimation based on partition recovery). In the context of Example 5.1, based on (5.1) we could estimate the parameters a and b directly from the entries of $\boldsymbol{S}(T)$ – or rather its estimated version $\widehat{\boldsymbol{\Sigma}}(T)$ – instead of using its eigenvalues. However, a main obstacle that we need to address in this setting is that the relation between $\boldsymbol{\Omega}$ and the entries of $\boldsymbol{S}(T)$ depends on the group labels. This implies that we would have to first recover the latent partition before we can estimate the parameters based on these relations. Although this is a reasonable procedure in practice, the recovery performance will depend strongly on how good the partition recovery is in the first place.

6. Theoretical analysis. In the following we provide the theoretical analysis for the spectral inference algorithms (Algorithms 4.1 and 5.1) discussed above. We first prove the concentration result in Theorem 4.1, which underlies the success of our spectral algorithms. The proof of this result follows from a series of lemmas that we establish in Subsections 6.1.1 and 6.1.2. Afterwards we show in Subsection 6.2 how Theorem 4.1 implies the partition recovery performance guarantees stated in Theorem 4.2, and prove Theorem 5.2 in Subsection 6.3.

6.1. Proof of Theorem 4.1: Concentration of the sample covariance matrix. In the next two subsections we prove Theorem 4.1 through a series of lemmas. The proof uses the triangle inequality

$$\|\widehat{\boldsymbol{\Sigma}}(T) - \boldsymbol{S}(T)\| \leq \|\widehat{\boldsymbol{\Sigma}}(T) - \boldsymbol{\Sigma}(T)\| + \|\boldsymbol{\Sigma}(T) - \boldsymbol{S}(T)\|, \tag{6.1}$$

and separate bounds for the two terms on the right-hand side. Then, using the matrix concentration results in Lemma 2.3, the proof of Theorem 4.1 follows. More specifically, the result follows from Corollary 6.5 in Subsection 6.1.2 and Lemma 6.1 in Subsection 6.1.1.

6.1.1. Bounding the first term in (6.1). We now establish a bound for the distance $\|\widehat{\boldsymbol{\Sigma}}(T) - \boldsymbol{\Sigma}(T)\|$ between the true and the sample covariances.

Lemma 6.1. *Under the conditions in Theorem 4.1, the following bound holds with probability* $1 - 5s^{-1} - \epsilon$*:*

$$\left\|\widehat{\boldsymbol{\Sigma}}(T) - \boldsymbol{\Sigma}(T)\right\| \leq B(T), \tag{6.2}$$

where $B(T)$ *is as defined in Theorem 4.1.*

For the proof of Lemma 6.1 we make use of the following result established in [3].

Lemma 6.2. *[3, Theorem 2.1] Let* $\boldsymbol{x} \in \mathbb{R}^n$ *be a zero-mean sub-Gaussian vector satisfying Assumption 3.3, with covariance matrix* $\boldsymbol{\Sigma}$*. Let* $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(i)}$ *be i.i.d samples drawn from* $\boldsymbol{x}$*. Consider the sample covariance matrix* $\widehat{\boldsymbol{\Sigma}} := \frac{1}{s}\sum_{i=1}^{s}[\boldsymbol{x}^{(i)} - \bar{\boldsymbol{x}}][\boldsymbol{x}^{(i)} - \bar{\boldsymbol{x}}]^\top$*, where* $\bar{\boldsymbol{x}} := \frac{1}{s}\sum_{i=1}^{s}\boldsymbol{x}^{(i)}$ *is the sample mean. Then, with probability at least* $1 - 5s^{-1}$ *we have*

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F \leq c \cdot \|\boldsymbol{\Sigma}\| \cdot r_e(\boldsymbol{\Sigma}) \cdot \sqrt{\frac{\ln s}{s}}, \tag{6.3}$$

where $r_e(\boldsymbol{\Sigma}) = \frac{\sum_i \lambda_i(\boldsymbol{\Sigma})}{\|\Sigma\|}$ *is the effective rank of* $\boldsymbol{\Sigma}$*,* c *is a constant that depends on* c_0 *in Assumption 3.3, and* $\|\cdot\|_F$ *denotes the Frobenius norm.*

Notice that we always have $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \leq \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F$ and, for the scenario considered here, $\|\boldsymbol{\Sigma}(T)\| = 1$, as by construction the dominant eigenvalue of $\boldsymbol{L}$ equals 1. Lemma 6.2 thus provides a useful tool to prove Lemma 6.1, provided that we can find an upper bound on the effective rank $r_e(\boldsymbol{\Sigma}(T))$.

Lemma 6.3 (Bound on the effective rank). *Under the conditions stated in Theorem 4.1, with probability* $1 - \epsilon$ *the effective rank satisfies the bound*

$$r_e(\boldsymbol{\Sigma}(T)) \leq 1 + \sum_{i=2}^{k}(|\mu_i| + M)^{2T} + (n-k)M^{2T}. \tag{6.4}$$

Proof. Recall that we denote the eigenvalues of $\boldsymbol{L}$ by λ_i and the eigenvalues of $\boldsymbol{\mathcal{L}}$ by μ_i. We assume, without loss of generality, that the eigenvalues have been sorted in descending order according to their magnitude, i.e., $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$ and $|\mu_1| \geq |\mu_2| \geq \ldots \geq |\mu_n|$,

respectively. From the definition of the effective rank and the fact that $\|\boldsymbol{\Sigma}(T)\| = 1$ it follows that

$$r_e(\boldsymbol{\Sigma}(T)) = \sum_{i=1}^{n} \lambda_i(\boldsymbol{\Sigma}(T)) = \sum_{i=1}^{n} \lambda_i^{2T} = \sum_{i=1}^{n} |\lambda_i|^{2T}.$$

We can decompose this sum into three parts as follows

$$r_e(\boldsymbol{\Sigma}(T)) = 1 + \sum_{i=2}^{k} |\lambda_i|^{2T} + \sum_{j=k+1}^{n} |\lambda_j|^{2T}.$$

In order to relate the (random) eigenvalues λ_i to quantities that can be expressed in terms of the model parameters, we make use of the triangle inequality $|\lambda_i| \leq |\mu_i| + |\lambda_i - \mu_i|$. Using the fact that $\mu_i = 0$ for $i > k$, we obtain the following bound with probability at least $1 - \epsilon$,

$$r_e(\boldsymbol{\Sigma}(T)) \leq 1 + \sum_{i=2}^{k} (|\mu_i| + M)^{2T} + \sum_{j=k+1}^{n} M^{2T} = 1 + \sum_{i=2}^{k} (|\mu_i| + M)^{2T} + (n-k)M^{2T},$$

where the inequality follows from Weyl's Theorem (see e.g., [13]), which states that $|\mu_i - \lambda_i| \leq \|\boldsymbol{L} - \boldsymbol{\mathcal{L}}\|$ for all i, and Lemma 2.3 implies that $\|\boldsymbol{L} - \boldsymbol{\mathcal{L}}\| \leq M$ with probability $1 - \epsilon$. ∎

Proof of Lemma 6.1. By combining the results from Lemmas 6.2 and 6.3, Lemma 6.1 follows. ∎

6.1.2. Bounding the second term in (6.1). We derive a bound for the term $\|\boldsymbol{\Sigma}(T) - \boldsymbol{S}(T)\|$, limiting the difference between the true and the population covariances. We start by establishing the following lemma.

Lemma 6.4. *Let $\boldsymbol{L}$ be the normalized adjacency matrix corresponding to a graph drawn from an SBM. Then for any integer $t > 0$, we have that*

$$\|\boldsymbol{L}^t - \boldsymbol{\mathcal{L}}^t\| \leq t\|\boldsymbol{L} - \boldsymbol{\mathcal{L}}\|. \tag{6.5}$$

Proof. We will show that for any integer $t > 0$

$$\|\boldsymbol{L}^t - \boldsymbol{\mathcal{L}}^t\| \leq \|\boldsymbol{L} - \boldsymbol{\mathcal{L}}\| + \|\boldsymbol{L}^{t-1} - \boldsymbol{\mathcal{L}}^{t-1}\|. \tag{6.6}$$

Equation (6.5) then directly follows from repeated application of (6.6). To show (6.6), notice that

$$\begin{aligned} \left\|\boldsymbol{L}^t - \boldsymbol{\mathcal{L}}^t\right\| &= \left\|(\boldsymbol{L} - \boldsymbol{\mathcal{L}})\,\boldsymbol{L}^{t-1} + \boldsymbol{\mathcal{L}}\left(\boldsymbol{L}^{t-1} - \boldsymbol{\mathcal{L}}^{t-1}\right)\right\| \\ &\leq \|\boldsymbol{L} - \boldsymbol{\mathcal{L}}\|\,\|\boldsymbol{L}^{t-1}\| + \|\boldsymbol{\mathcal{L}}\|\left\|\boldsymbol{L}^{t-1} - \boldsymbol{\mathcal{L}}^{t-1}\right\| = \|\boldsymbol{L} - \boldsymbol{\mathcal{L}}\| + \left\|\boldsymbol{L}^{t-1} - \boldsymbol{\mathcal{L}}^{t-1}\right\|, \end{aligned}$$

where the last equality follows from the fact that $\|\boldsymbol{\mathcal{L}}\| = \|\boldsymbol{L}\| = 1$. ∎

Lemma 6.4 directly provides us with the following corollary, which establishes the desired bound for $\|\boldsymbol{\Sigma}(T) - \boldsymbol{S}(T)\|$ needed in Theorem 4.1.

Corollary 6.5. *The following bound holds between the covariances $\boldsymbol{\Sigma}(T)$ and $\boldsymbol{S}(T)$:*

$$\|\boldsymbol{\Sigma}(T) - \boldsymbol{S}(T)\| \leq 2T\|\boldsymbol{L} - \boldsymbol{\mathcal{L}}\| \leq 2TM, \tag{6.7}$$

where the last inequality holds with probability $1 - \epsilon$.

6.2. Implications for partition recovery. In the following we explain how the above concentration results translate into guarantees for the partition recovery problem. Subsection 6.2.1 provides intermediate results on the spectral geometry of the problem and discusses the precise definition of misclassified nodes. The proof of Theorem 4.2 is then given in Subsection 6.2.2.

6.2.1. Partition recovery problem. In order to characterize the recovery performance, we first clarify what we mean by a node being clustered correctly. To do this we follow the literature on SBMs [26, 27]. Recall that within our partition recovery algorithm, we perform a k-means clustering procedure on the rows of the matrix $\widetilde{\boldsymbol{W}}$ of (row-normalized) dominant eigenvectors associated with the estimated covariance $\widehat{\boldsymbol{\Sigma}}(T)$. This procedure assigns each row i to a centroid vector $\boldsymbol{k}_i$ corresponding to a cluster.

Let us denote the 'true' centroid vector that a node i is assigned to according to the expected population covariance $\boldsymbol{S}(T)$ by $\boldsymbol{\kappa}_i$, where $i = 1, \dots, n$. From the definition of $\boldsymbol{S}(T)$ it follows that each $\boldsymbol{\kappa}_i^\top$ corresponds to a row of the matrix $\widetilde{\boldsymbol{V}}$ of the (row-normalized) top k eigenvectors of $\boldsymbol{\mathcal{L}}$. Moreover, we know that there are exactly k different centroids. However, if some of the eigenvalues of $\boldsymbol{S}(T)$ are not unique, the eigenvector matrix $\widetilde{\boldsymbol{V}}$ will not be uniquely defined, even though the subspace spanned by these vectors will be unique. As we are interested in the distance between the vectors $\widetilde{\boldsymbol{V}}$ and their approximation via $\widetilde{\boldsymbol{W}}$, our definition of misclassification has to include an orthogonal matrix $\boldsymbol{\mathcal{Q}} \in \mathbb{R}^{k\times k}$ that minimizes the orthogonal Procrustes distance $\|\widetilde{\boldsymbol{W}} - \widetilde{\boldsymbol{V}}\boldsymbol{\mathcal{Q}}\|_F$ (see also [26, 27]). In particular, this minimality property of $\boldsymbol{\mathcal{Q}}$ will be essential for Lemma 6.9.

Definition 6.6 (Set of misclassified nodes). *The set of misclassified nodes* $\mathbb{M}$ *is defined as*

$$\mathbb{M} := \{i : \|\boldsymbol{k}_i^\top - \boldsymbol{\kappa}_i^\top \boldsymbol{\mathcal{Q}}\| > \|\boldsymbol{k}_i^\top - \boldsymbol{\kappa}_j^\top \boldsymbol{\mathcal{Q}}\| \text{ for some } j \neq i\}. \tag{6.8}$$

In words, this means that we call a node i misclassified if the centroid $\boldsymbol{k}_i$ to which it has been assigned is closer to some other population centroid $\boldsymbol{\kappa}_j$ than it is to its true population centroid $\boldsymbol{\kappa}_i$.

By accounting for the special geometry of the eigenvectors $\widetilde{\boldsymbol{V}}$, we can establish the following necessary condition for a node to be misclassified.

Lemma 6.7. *If a node i is misclassified, then $\|\boldsymbol{k}_i^\top - \boldsymbol{\kappa}_i^\top \boldsymbol{\mathcal{Q}}\| \geq 1/\sqrt{2}$.*

Proof. First, we know from Lemma 2.1 that $\widetilde{\boldsymbol{V}} = \boldsymbol{G}\boldsymbol{\mathcal{U}}$ for some orthogonal matrix $\boldsymbol{\mathcal{U}}$. Hence, all population centroids $\boldsymbol{\kappa}_i$ are of unit length and orthogonal to each other. In particular, for all j and i belonging to different blocks, we have

$$\|\boldsymbol{\kappa}_i^\top - \boldsymbol{\kappa}_j^\top\| = \|(\boldsymbol{g}_i^\top - \boldsymbol{g}_j^\top)\boldsymbol{\mathcal{U}}\| = \sqrt{2}, \tag{6.9}$$

where $\boldsymbol{g}_i$ is the true population cluster indicator vector corresponding to node i (see Lemma 2.1).

Second, by the triangle inequality and the orthogonality of $\boldsymbol{\mathcal{Q}}$, we know that for all j and i in different blocks, we have

$$\|\boldsymbol{\kappa}_i^\top - \boldsymbol{\kappa}_j^\top\| \leq \|\boldsymbol{\kappa}_i^\top \boldsymbol{\mathcal{Q}} - \boldsymbol{k}_i^\top\| + \|\boldsymbol{k}_i^\top - \boldsymbol{\kappa}_j^\top \boldsymbol{\mathcal{Q}}\|. \tag{6.10}$$

Plugging in the result from (6.9) into (6.10), we obtain that $\|\boldsymbol{k}_i^\top - \boldsymbol{\kappa}_j^\top \boldsymbol{\mathcal{Q}}\| \geq \sqrt{2} - \|\boldsymbol{k}_i^\top - \boldsymbol{\kappa}_i^\top \boldsymbol{\mathcal{Q}}\|$. From Definition 6.6, for a misclassified node i there exists some j such that

$$\|\boldsymbol{k}_i^\top - \boldsymbol{\kappa}_i^\top \boldsymbol{\mathcal{Q}}\| > \|\boldsymbol{k}_i^\top - \boldsymbol{\kappa}_j^\top \boldsymbol{\mathcal{Q}}\| \geq \sqrt{2} - \|\boldsymbol{k}_i^\top - \boldsymbol{\kappa}_i \boldsymbol{\mathcal{Q}}\|,$$

where the last inequality is true by our argument above. The result follows after some simple algebraic manipulations. ∎

To arrive at our result for the misclassification rate in Theorem 4.2, we further follow the literature of SBMs [1, 26, 27], and establish the following two lemmas (Lemmas 6.8 and 6.9). The first lemma provides us with a relationship between the difference of the cluster centroids obtained from the k-means clustering and the true population cluster on the one side, and the difference between the eigenvectors of $\widehat{\boldsymbol{\Sigma}}(T)$ and $\boldsymbol{S}(T)$ on the other side.

Lemma 6.8. *Denote the matrix of cluster centroids obtained by performing k-means on the rows of $\widetilde{\boldsymbol{W}}$ as $\boldsymbol{K} := [\boldsymbol{k}_1, \dots, \boldsymbol{k}_n]^\top \in \mathbb{R}^{n\times k}$. Then,*

$$\|\boldsymbol{K} - \widetilde{\boldsymbol{V}}\boldsymbol{\mathcal{Q}}\|_F \leq 2\|\widetilde{\boldsymbol{W}} - \widetilde{\boldsymbol{V}}\boldsymbol{\mathcal{Q}}\|_F. \tag{6.11}$$

Proof. By the triangle inequality,

$$\|\boldsymbol{K} - \widetilde{\boldsymbol{V}}\boldsymbol{\mathcal{Q}}\|_F \leq \|\boldsymbol{K} - \widetilde{\boldsymbol{W}}\|_F + \|\widetilde{\boldsymbol{W}} - \widetilde{\boldsymbol{V}}\boldsymbol{\mathcal{Q}}\|_F.$$

From the definition of k-means, however, we know that $\|\widetilde{\boldsymbol{W}} - \boldsymbol{K}\|_F \leq \|\widetilde{\boldsymbol{W}} - \widetilde{\boldsymbol{V}}\boldsymbol{\mathcal{Q}}\|_F$ as $\boldsymbol{K}$ is chosen in order to minimize the distances of the centroids to the individual rows. Hence, the claim follows. ∎

The second lemma builds on the Davis-Kahan theorem [7, 44], and bounds the difference between the row-normalized eigenvectors of $\widehat{\boldsymbol{\Sigma}}(T)$ and $\boldsymbol{S}(T)$.

Lemma 6.9 (Davis-Kahan). *Following the definitions in Theorem 4.2, we have*

$$\|\widetilde{\boldsymbol{W}} - \widetilde{\boldsymbol{V}}\boldsymbol{\mathcal{Q}}\|_F \leq \frac{\sqrt{8k}\|\widehat{\boldsymbol{\Sigma}}(T) - \boldsymbol{S}(T)\|}{\tau\xi_k}. \tag{6.12}$$

Proof. The fact that

$$\|\boldsymbol{W} - \boldsymbol{V}\boldsymbol{\mathcal{Q}}\|_F \leq \frac{\sqrt{8k}\|\widehat{\boldsymbol{\Sigma}}(T) - \boldsymbol{S}(T)\|}{\xi_k},$$

holds for the unnormalized versions of $\boldsymbol{W}$ and $\boldsymbol{V}$ is a restatement of the Davis-Kahan Theorem as can be found in, e.g., [44, Theorem 2].

By definition, we have $\widetilde{\boldsymbol{V}} = \operatorname{diag}(\boldsymbol{V}\boldsymbol{V}^\top)^{-1/2}\boldsymbol{V}$ and $\widetilde{\boldsymbol{W}} = \operatorname{diag}(\boldsymbol{W}\boldsymbol{W}^\top)^{-1/2}\boldsymbol{W}$. Let us define $\tau_i := \min\{[\boldsymbol{V}\boldsymbol{V}^\top]_{ii}^{1/2}, [\boldsymbol{W}\boldsymbol{W}^\top]_{ii}^{1/2}\}$, which is the smaller of the 2-norms of the i-th rows of $\boldsymbol{V}$ and $\boldsymbol{W}$, respectively. Based on these definitions we see that $\tau = \min_i \tau_i$. Now, as row-normalization corresponds to diagonal scaling we obtain

$$\|\widetilde{\boldsymbol{W}} - \widetilde{\boldsymbol{V}}\boldsymbol{\mathcal{Q}}\|_F \leq \frac{1}{\tau}\|\boldsymbol{W} - \boldsymbol{V}\boldsymbol{\mathcal{Q}}\|_F \leq \frac{\sqrt{8k}\|\widehat{\boldsymbol{\Sigma}}(T) - \boldsymbol{S}(T)\|}{\tau\xi_k}.$$

∎

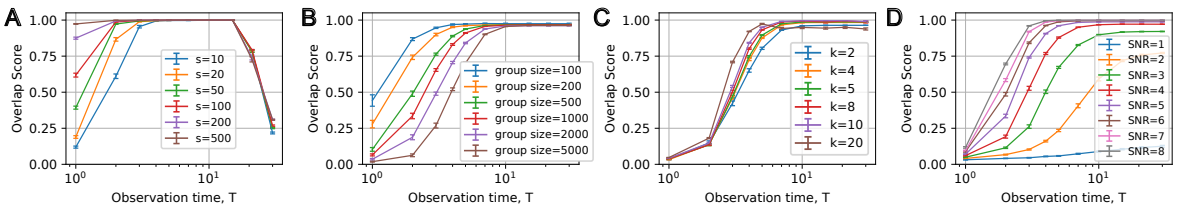

Figure 1. ***Partition recovery performance for SBM inference.*** ***A-D.*** *The recovery performance of Algorithm 4.1 is plotted against the observation time at which the samples are drawn. Each point in the plot corresponds to an average over 30 experiments, the error bars denote the standard error of the mean. Within Algorithm 4.1 we run k-means 10 times and pick the best result in terms of the k-means objective function as our final partition.* ***A.*** *Recovery performance for varying number of samples for the SBM discussed in the text.* ***B.*** *Recovery performance in a $k = 2$ group planted partition model (see Example 5.1) for varying group sizes ($SNR = 5, s = 50, \rho = 30/n$).* ***C.*** *Recovery performance in a k-group planted partition model (see Example 5.1) for varying number of groups ($n = 5000, SNR = 5, s = 50, \rho = 30/n$).* ***D.*** *Recovery performance in a $k = 5$ group planted partition model (see Example 5.1) for different SNR ($n = 2000, s = 50, \rho = 30/n$).*

6.2.2. Proof of Theorem 4.2. Based on the results established thus far, we can provide a proof of the bound of the misclassification rate $q := |\mathbb{M}|/n$ as stated in Theorem 4.2.

Proof. We define the set $\mathbb{M}_2 := \{i : \|\boldsymbol{k}_i^\top - \boldsymbol{\kappa}_i^\top \boldsymbol{\mathcal{Q}}\| \geq 1/\sqrt{2}\} \supset \mathbb{M}$, which includes all misclassified nodes according to Lemma 6.7. From Lemmas 6.8 and 6.9, and Theorem 4.1, it follows that

$$q = \frac{|\mathbb{M}|}{n} \leq \frac{|\mathbb{M}_2|}{n} = \frac{1}{n}\sum_{i\in\mathbb{M}_2} 1 \leq \frac{2}{n}\sum_{i\in\mathbb{M}_2} \|\boldsymbol{k}_i^\top - \boldsymbol{\kappa}_i^\top \boldsymbol{\mathcal{Q}}\|^2 = \frac{2}{n}\|\boldsymbol{K} - \widetilde{\boldsymbol{V}}\boldsymbol{\mathcal{Q}}\|_F^2$$
$$\leq \frac{8}{n}\|\widetilde{\boldsymbol{W}} - \widetilde{\boldsymbol{V}}\boldsymbol{\mathcal{Q}}\|_F^2 \leq \frac{8}{n}\left(\frac{\sqrt{8k}\|\widehat{\boldsymbol{\Sigma}}(T) - \boldsymbol{S}(T)\|}{\tau\xi_k}\right)^2.$$

■

6.3. Implications for parameter estimation. Here we show how our concentration results allow us to prove Theorem 5.2.

Proof. Note that we can write the relative estimation error as:

$$\frac{|\hat{a} - a|}{a} = \frac{|\lambda_2(\widehat{\boldsymbol{\Sigma}}(T))^{1/2T}\rho n - \lambda_2(\boldsymbol{S}(T))^{1/2T}\rho n|}{\lambda_2(\boldsymbol{S}(T))^{1/2T}\rho n + \rho n} = \frac{|\lambda_2(\widehat{\boldsymbol{\Sigma}}(T))^{1/2T} - \lambda_2(\boldsymbol{S}(T))^{1/2T}|}{\lambda_2(\boldsymbol{S}(T))^{1/2T} + 1}$$
$$\leq \frac{|\lambda_2(\widehat{\boldsymbol{\Sigma}}(T)) - \lambda_2(\boldsymbol{S}(T))|^{1/2T}}{\lambda_2(\boldsymbol{\mathcal{L}}) + 1} \leq \frac{\|\boldsymbol{\Sigma}(T) - \boldsymbol{S}(T)\|^{1/2T}}{\lambda_2(\boldsymbol{\mathcal{L}}) + 1}.$$

The first inequality follows from Jensen's inequality since the function $f(x) = x^{1/2T}$ is concave. The second inequality is a direct application of Weyl's inequality. Hence, using Theorem 4.1 it follows that with probability $1 - 5s^{-1} - \epsilon$,

$$\eta \leq \frac{(2TM + B(T))^{1/2T}}{\lambda_2(\boldsymbol{\mathcal{L}}) + 1}.$$

■

7. Numerical experiments. We present some computational experiments using our algorithms for partition recovery and parameter estimation to demonstrate their performance in practice.

7.1. Partition recovery. To gain intuition about the qualitative features of the recovery problem, we first consider the following example SBM consisting of 5 groups with varying sizes $\boldsymbol{n}_g = [200, 300, 200, 300, 280]^\top$, and affinity matrix

$$\boldsymbol{\Omega} = \begin{bmatrix} 0.5 & 0.3 & 0.2 & 0.1 & 0.2 \\ 0.3 & 0.5 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.4 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.2 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.1 & 0.2 & 0.4 \end{bmatrix}.$$

The partition recovery performance is shown in Figure 1A for a varying number of samples s, as a function of the observation time T. Here, the recovery performance is measured via the overlap score Z defined as:

$$Z = \frac{z_{\text{actual}} - z_{\text{chance}}}{1 - z_{\text{chance}}}, \tag{7.1}$$

where z_{chance} is the probability of guessing the group assignments correctly and z_{actual} is the fraction of nodes that are correctly classified. The guessing probability is computed as $z_{\text{chance}} = n/\max(\boldsymbol{n}_g)$, where $\max(\boldsymbol{n}_g)$ is the size of the largest group. Note that z_{chance} is simply $\frac{1}{\#\text{groups}}$ for groups of equal size. The actual overlap z_{actual} is computed by solving an optimal assignment problem between the inferred group labels and the planted partition.

Overall, we can infer the partitions remarkably well using our simple spectral inference strategy, even though we never observe the actual network. The results in Figure 1A display a couple of noteworthy features consistent with our technical results, which are worth examining. Specifically, we see that there is a pronounced effect of the observation time: both for small as well as very large T, the performance is markedly worse than for intermediate times. Thus, if we have a fixed number of samples at our disposal, choosing the 'right' observation time can yield an improved recovery performance. Obtaining more samples from our dynamics, in contrast, has a more moderate effect and is most beneficial for small times.

At an intuitive level these results can be explained by the role that the network plays in terms of a linear filter. On the one hand, for short observation times, the original white noise input is only filtered weakly and thus contains still a large noise component, which is detrimental to recovery performance. On the other hand, as the output approaches its stationary state, the information that the signal carries about the network structure is completely washed out. For intermediate times, however, there is a trade-off between attenuating the noise within the observed signal, and not losing too much information about the partition structure. Theorems 4.1 and 4.2 quantify precisely how this trade-off is governed by the spectral features of the network.

Using a planted partition model with an affinity matrix of the form $\boldsymbol{\Omega} = \frac{(a-b)}{n}\boldsymbol{I}_k + \frac{b}{n}\mathbf{1}_k\mathbf{1}_k^\top$ (see Example 5.1), we further investigate numerically how the partition recovery performance is affected by the group sizes, the number of groups and the signal-to-noise ratio (SNR) within

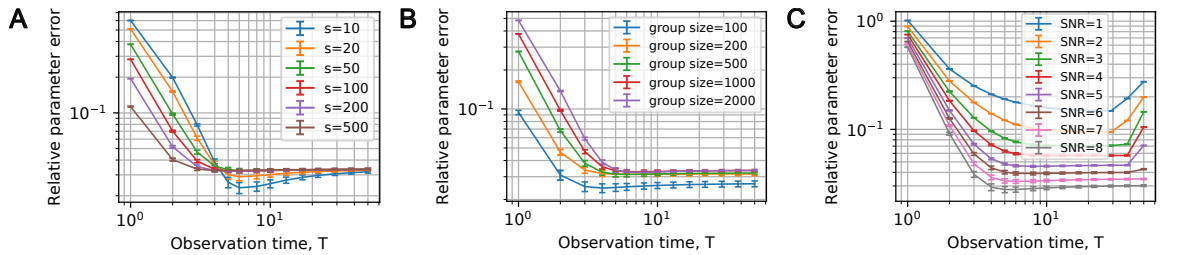

Figure 2. ***Parameter recovery performance for SBM inference.*** ***A-C.*** *The recovery performance of Algorithm 5.1 in terms of the relative error $\eta = |a - \hat{a}|/a$ is plotted against the observation time at which the samples are drawn. Each point corresponds to an average over 30 experiments and the error bars denote the standard error of the mean.* ***A.*** *Parameter recovery performance for varying number of samples for the SBM with $k = 2$ groups as discussed in the text ($n = 2000$, $SNR = 7$, $s = 50$, $\rho = 30/n$).* ***B.*** *Parameter recovery performance in a $k = 2$ group planted partition model (see Example 5.1) for varying group sizes ($SNR = 7$, $s = 50$, $\rho = 30/n$).* ***C.*** *Parameter recovery performance in a $k = 2$ group planted partition model (see Example 5.1) for different SNR ($n = 2000$, $s = 50$, $\rho = 30/n$).*

the underlying stochastic block model (see Figure 1B-D). Here the SNR is defined as [1],

$$\text{SNR} = \frac{(a-b)^2}{ka + k(k-1)b}, \tag{7.2}$$

so that $\text{SNR} = 1$ corresponds to the well-known detectability limit in the sparse regime of the SBM, for asymptotically large networks with $n \to \infty$ (see [1] for a detailed review).

Figure 1B shows that for a fixed number of groups (here $k = 2$) increasing the group sizes results in a shift of the recovery performance in terms of the observation time. However, if we fix the number of nodes and let the number of groups vary (Figure 1C) no such effect is observed. This hints at the fact that the number of nodes n (and thus the number of nodes per group) is one of the driving factors determining the difficulty in the detection of the SBM partition. Finally, in Figure 1D we plot the influence of the SNR inherent to the latent SBM on the detection performance. As expected, a reduced SNR results in a dramatically decreasing detection performance. Remarkably, for $\text{SNR} = 4$, using only $s = 50$ samples an almost perfect partition recovery is possible for a suitable observation time. We remark that we can reach here a nonzero recovery performance for larger times T even for $\text{SNR} = 1$ due to finite size effects.

7.2. Model parameter estimation. We now demonstrate numerically that we can infer the model parameters of the SBM using only the observed outputs. For simplicity, we focus on the case of the planted partition model with $k = 2$ groups of equal size, and assume that the average expected network density ρ, or equivalently the average expected mean degree $d_{\text{av}} = \rho n$, is known a priori (see also the discussion in the context of Example 5.1 and Subsection 5.2). We focus on the eigenvalue-based parameter estimation strategy using the estimator for the parameter a as discussed in Subsection 5.2.

Figure 2 shows the results of applying this parameter estimation scheme in numerical experiments. To quantify the parameter recovery performance we compute the relative recovery error $\eta = |a - \hat{a}|/a$. Similar to the partition recovery, we see in Figure 2A that the observation

time has a strong effect on the recovery performance and we see a marked decrease in the error from more than 10% initially, to around 3% for longer observation times.

Figure 2B-C demonstrates the effect of the group size and the SNR on the recovery performance. Similar to the partition recovery problem, we observe that smaller group sizes (a smaller network) and a larger SNR lead to better performance.

8. Discussion. Graph-based tools have become prevalent for the analysis of a range of different systems across the sciences. Network inference, the problem of determining the topology of a network based on a set of nodal observations, has accordingly become a paramount tool to enable this type of analysis. In this work we have advocated a fresh look onto this problem for applications in which exact inference is impossible, e.g., due to sampling constraints. Instead of inferring the exact network in terms of its adjacency matrix, we infer a generative model of the network structure, thereby learning a probability distribution over networks. Importantly, for a low-dimensional statistical model such as the SBM discussed here, this procedure can be implemented using far fewer samples than would be required to infer the exact network structure.

The proposed method is especially relevant for scenarios in which we aim to analyze the network further from a statistical perspective, e.g., if our ultimate goal is to perform community detection on the network. An added benefit of having fitted a generative model is that we can employ the model for further tasks afterwards, e.g., to generate surrogate network data.

Our work opens up several interesting avenues for future research. This includes the analysis of the inference of more realistic network models based on a setup as discussed here, ranging from degree-corrected SBMs [16] all the way to graphons [2, 19]. Likewise, it will be interesting to extend the above results to more general dynamical processes, including nonlinear dynamics. Another appealing line of future work would be to consider scenarios in which only a part of the network is observed [23], or additional (control) inputs are possible.

Appendix A. Omitted proofs.

A.1. Proof of Lemma 2.1. In order to prove Lemma 2.1, we provide a slightly more general discussion about the relationship between the eigenvectors of $\boldsymbol{\mathcal{L}}$ and the partition indicator matrix $\boldsymbol{G}$. Lemma 2.1 then follows directly from Corollary A.4 below.

Definition A.1. *(Normalized affinity matrix) Consider an SBM with affinity matrix $\boldsymbol{\Omega}$ and partition indicator matrix $\boldsymbol{G}$. Let us denote the vector of group sizes by $\boldsymbol{n}_g = \boldsymbol{G}^\top \mathbf{1}$ and define $\boldsymbol{N}_g := \operatorname{diag}(\boldsymbol{n}_g) = \boldsymbol{G}^\top \boldsymbol{G}$. We then define the normalized affinity matrix as*

$$\boldsymbol{\Gamma} = \boldsymbol{N}_g^{1/2} \operatorname{diag}(\boldsymbol{\Omega}\boldsymbol{n}_g)^{-1/2} \boldsymbol{\Omega} \operatorname{diag}(\boldsymbol{\Omega}\boldsymbol{n}_g)^{-1/2} \boldsymbol{N}_g^{1/2}.$$

Note that $\boldsymbol{\mathcal{L}}$ can be written as

$$\boldsymbol{\mathcal{L}} = \boldsymbol{G}\boldsymbol{N}_g^{-1/2} \boldsymbol{\Gamma} \boldsymbol{N}_g^{-1/2} \boldsymbol{G}^\top. \tag{A.1}$$

Lemma A.2. *Let $\boldsymbol{\mathcal{L}}$ be the expected normalized adjacency matrix of an SBM with a corresponding normalized affinity matrix $\boldsymbol{\Gamma}$ with eigenvalues λ_i and corresponding eigenvectors $\boldsymbol{u}_i$. Then, any eigenvector $\boldsymbol{v}_i$ of $\boldsymbol{\Gamma}$ will give rise to an eigenvector $\boldsymbol{v}_i = \boldsymbol{G}\boldsymbol{N}_g^{-1/2}\boldsymbol{u}_i$ of $\boldsymbol{\mathcal{L}}$ with the same eigenvalue.*

Proof. A simple computation shows that

$$\boldsymbol{\mathcal{L}}\boldsymbol{G}\boldsymbol{N}_g^{-1/2}\boldsymbol{u}_i = \boldsymbol{G}\boldsymbol{N}_g^{-1/2}\boldsymbol{\Gamma}\boldsymbol{N}_g^{-1/2}\boldsymbol{G}^\top\boldsymbol{G}\boldsymbol{N}_g^{-1/2}\boldsymbol{u}_i = \boldsymbol{G}\boldsymbol{N}_g^{-1/2}\boldsymbol{\Gamma}\boldsymbol{u}_i = \lambda_i\boldsymbol{G}\boldsymbol{N}_g^{-1/2}\boldsymbol{u}_i = \lambda_i\boldsymbol{v}_i \quad \blacksquare$$

Corollary A.3. *The partition indicator matrix $\boldsymbol{G}$ can be written as a linear combination of the eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ corresponding to the k nonzero eigenvectors of $\boldsymbol{\mathcal{L}}$.*

Proof. From the previous lemma we know that the eigenvectors of $\boldsymbol{\mathcal{L}}$ can be written as $\boldsymbol{v}_i = \boldsymbol{G}\boldsymbol{N}_g^{-1/2}\boldsymbol{u}_i$, where $\boldsymbol{u}_i$ are the eigenvectors of $\boldsymbol{\Gamma}$. Assembling these relationships into matrices we obtain

$$\boldsymbol{V} = \boldsymbol{G}\boldsymbol{N}_g^{-1/2}\boldsymbol{U}. \tag{A.2}$$

Since $\boldsymbol{U}$ is an orthogonal matrix and $\boldsymbol{N}_g^{1/2}$ is diagonal, both matrices are invertible. It then follows that

$$\boldsymbol{G} = \boldsymbol{V}\boldsymbol{U}^\top\boldsymbol{N}_g^{1/2}. \tag{A.3}$$ $\blacksquare$

Corollary A.4. *Consider the row-normalized eigenvector matrix $\widetilde{\boldsymbol{V}} = \operatorname{diag}(\boldsymbol{V}\boldsymbol{V}^\top)^{-1/2}\boldsymbol{V}$. There exists an orthogonal transformation $\boldsymbol{\mathcal{U}}$ such that $\widetilde{\boldsymbol{V}} = \boldsymbol{G}\boldsymbol{\mathcal{U}}$.*

Proof. Note that

$$\operatorname{diag}(\boldsymbol{V}\boldsymbol{V}^\top) = \operatorname{diag}(\boldsymbol{G}\boldsymbol{N}_g^{-1/2}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{N}_g^{-1/2}\boldsymbol{G}^\top) = \operatorname{diag}(\boldsymbol{G}\boldsymbol{N}_g^{-1}\boldsymbol{G}^\top).$$

Combining this equation with (A.2) and the fact that $\boldsymbol{G}$ is an indicator matrix it follows that

$$\widetilde{\boldsymbol{V}} = \operatorname{diag}(\boldsymbol{G}\boldsymbol{N}_g^{-1}\boldsymbol{G}^\top)^{-1/2}\boldsymbol{V} = \operatorname{diag}(\boldsymbol{G}\boldsymbol{N}_g^{-1}\boldsymbol{G}^\top)^{-1/2}\boldsymbol{G}\boldsymbol{N}_g^{-1/2}\boldsymbol{U} = \boldsymbol{G}\boldsymbol{N}_g^{1/2}\boldsymbol{N}_g^{-1/2}\boldsymbol{U} = \boldsymbol{G}\boldsymbol{U}.$$

Since $\boldsymbol{U}$ corresponds to the eigenvector basis of the symmetric normalized affinity matrix $\boldsymbol{\Gamma}$, we know that $\boldsymbol{U}$ is orthogonal. $\blacksquare$

A.2. Proof of Lemma 2.3.

Proof. For notational convenience we define the matrices $\boldsymbol{\mathcal{D}} := \mathbb{E}[\boldsymbol{D}]$ and $\boldsymbol{H} := \boldsymbol{\mathcal{D}}^{-1/2}\boldsymbol{A}\boldsymbol{\mathcal{D}}^{-1/2}$ From the triangle inequality we have

$$\|\boldsymbol{L} - \boldsymbol{\mathcal{L}}\| \leq \|\boldsymbol{L} - \boldsymbol{H}\| + \|\boldsymbol{H} - \boldsymbol{\mathcal{L}}\|. \tag{A.4}$$

To obtain our desired concentration inequality, we bound the two terms on the right-hand side of (A.4) separately.

For the first term we note that since $\|\boldsymbol{L}\| \leq 1$,

$$\begin{aligned}
\|\boldsymbol{L} - \boldsymbol{H}\| &= \|\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2} - \boldsymbol{\mathcal{D}}^{-1/2}\boldsymbol{A}\boldsymbol{\mathcal{D}}^{-1/2}\| = \|\boldsymbol{L} - \boldsymbol{\mathcal{D}}^{-1/2}\boldsymbol{D}^{1/2}\boldsymbol{L}\boldsymbol{D}^{1/2}\boldsymbol{\mathcal{D}}^{-1/2}\| \\
&= \left\|(\boldsymbol{I} - \boldsymbol{\mathcal{D}}^{-1/2}\boldsymbol{D}^{1/2})\boldsymbol{L}\boldsymbol{D}^{1/2}\boldsymbol{\mathcal{D}}^{-1/2} + \boldsymbol{L}(\boldsymbol{I} - \boldsymbol{D}^{1/2}\boldsymbol{\mathcal{D}}^{-1/2})\right\| \\
&\leq \|\boldsymbol{I} - \boldsymbol{\mathcal{D}}^{-1/2}\boldsymbol{D}^{1/2}\|\|\boldsymbol{D}^{1/2}\boldsymbol{\mathcal{D}}^{-1/2}\| + \|\boldsymbol{I} - \boldsymbol{\mathcal{D}}^{-1/2}\boldsymbol{D}^{1/2}\|. \qquad \text{(A.5)}
\end{aligned}$$

We can now analyze the term $\|\boldsymbol{I} - \boldsymbol{\mathcal{D}}^{-1/2}\boldsymbol{D}^{1/2}\|$ as follows:

$$\|\boldsymbol{I} - \boldsymbol{\mathcal{D}}^{-1/2}\boldsymbol{D}^{1/2}\| = \max_i \left|1 - \sqrt{D_{ii}/\mathcal{D}_{ii}}\right| \leq \max_i |1 - D_{ii}/\mathcal{D}_{ii}|\,,$$

where we have used the fact that the spectral norm of a diagonal matrix is its maximal diagonal entry. To make use of the above result, we need to obtain a concentration result for the degrees. Recalling that $\delta_i = \mathbb{E}(d_i) = \mathcal{D}_{ii}$, we can use Chernoff's inequality to obtain a bound of the form:

$$\mathbb{P}(|d_i - \delta_i| \geq t\delta_i) \leq \frac{\epsilon}{2n} \quad \text{for } t \geq \sqrt{\frac{\log 4n/\epsilon}{\delta_i}}.$$

If we chose $t_0 = \sqrt{\log(4n/\epsilon)/\delta_{\min}}$, the above bound holds for all i, and by a simple transformation we obtain $\mathbb{P}(|d_i/\delta_i - 1| \geq t_0) \leq \epsilon/2n$ for all i. Using the union bound, we thus have with probability at least $1 - \epsilon/2$:

$$\|\boldsymbol{I} - \boldsymbol{\mathcal{D}}^{-1/2}\boldsymbol{D}^{1/2}\| = \max_i \left|1 - \sqrt{\frac{D_{ii}}{\mathcal{D}_{ii}}}\right| \leq \sqrt{\frac{\log(4n/\epsilon)}{\delta_{\min}}}.$$

Therefore, plugging this result into (A.5), with probability $1 - \epsilon/2$ it holds that

$$\|\boldsymbol{L} - \boldsymbol{H}\| \leq t_0(t_0 + 1) + t_0. \tag{A.6}$$

For the second term $\|\boldsymbol{H} - \boldsymbol{\mathcal{L}}\|$, we can use the following matrix concentration inequality.

Lemma A.5 (Bernstein Matrix inequality [5]). *Let $\boldsymbol{X}_i$ be independent $n \times n$ dimensional random Hermitian matrices such that $\|\boldsymbol{X}_i - \mathbb{E}[\boldsymbol{X}_i]\| \leq M \quad \forall i$, and $\boldsymbol{X} = \sum_i \boldsymbol{X}_i$. Then for any $a > 0$:*

$$\mathbb{P}(\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\| > a) \leq 2n \exp\left(\frac{-a^2}{2v^2 + 2Ma/3}\right),$$

where $v^2 = \|\sum_i \operatorname{var}(\boldsymbol{X}_i)\| = \|\sum_i \mathbb{E}[\boldsymbol{X}_i^2] - \mathbb{E}[\boldsymbol{X}_i]^2\|$.

To apply this result, note that we can decompose $[\boldsymbol{H} - \boldsymbol{\mathcal{L}}] = \sum_{i \leq j} \boldsymbol{X}_{ij}$, where $\boldsymbol{X}_{ij}$ are the Hermitian matrices:

$$\boldsymbol{X}_{ij} = \begin{cases} \frac{A_{ij} - p_{ij}}{\sqrt{\delta_i \delta_j}}(\boldsymbol{E}^{ij} + \boldsymbol{E}^{ij}), & \text{for } i \neq j, \\ \frac{A_{ii} - p_{ii}}{\delta_i}\boldsymbol{E}^{ii}, & \text{for } i = j. \end{cases}$$

Here, $\boldsymbol{E}^{ij}$ is the canonical basis matrix with entries $[\boldsymbol{E}^{ij}]_{ij} = 1$ and 0 otherwise. As $\mathbb{E}[\boldsymbol{X}_{ij}] = 0$ we compute $\operatorname{var}(\boldsymbol{X}_{ij}) = \mathbb{E}[\boldsymbol{X}_{ij}^2]$ as follows:

$$\operatorname{var}(\boldsymbol{X}_{ij}) = \begin{cases} \frac{1}{\delta_i \delta_j}(p_{ij} - p_{ij}^2)(\boldsymbol{E}^{ii} + \boldsymbol{E}^{jj}), & \text{for } i \neq j, \\ \frac{1}{\delta_i^2}(p_{ii} - p_{ii}^2)\boldsymbol{E}^{ii}, & \text{for } i = j. \end{cases}$$

From this we can bound v^2 as

$$v^2 = \left\| \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\delta_i \delta_j} (p_{ij} - p_{ij}^2) \boldsymbol{E}^{ii} \right\| = \max_i \left(\sum_{j=1}^{n} \frac{1}{\delta_i \delta_j} p_{ij} - \sum_{j=1}^{n} \frac{1}{\delta_i \delta_j} p_{ij}^2 \right)$$
$$\leq \max_i \left(\frac{1}{\delta_{\min}} \sum_{j=1}^{n} \frac{1}{\delta_i} p_{ij} \right) = \frac{1}{\delta_{\min}},$$

where we have used the bound $1/\delta_{\min} \geq 1/\delta_j$.

Based on the above calculations, we select $a = \sqrt{3 \log(4n/\epsilon)/\delta_{\min}} = \sqrt{3} t_0$, which leads to our desired bound for the second term

$$\mathbb{P}(\|\boldsymbol{H} - \boldsymbol{\mathcal{L}}\| > a) \leq 2n \exp\left(\frac{-a^2}{2/\delta_{\min} + 2a/(3\delta_{\min})} \right)$$
$$\leq 2n \exp\left(-\frac{3 \log(4n/\epsilon)}{3} \right) \leq \epsilon/2, \tag{A.7}$$

where the first inequality follows from the fact that $M \leq 1/\delta_{\min}$ (cf. Lemma A.5), and for the second inequality we used the fact that $\mathcal{M}(\boldsymbol{\mathcal{A}}) \in \mathcal{M}_n(\epsilon)$ to determine a lower bound on $\delta_{\min}$. Finally, by combining the results in (A.6) and (A.7) we obtain that with probability $1 - \epsilon$,

$$\|\boldsymbol{L} - \boldsymbol{\mathcal{L}}\| \leq \|\boldsymbol{L} - \boldsymbol{H}\| + \|\boldsymbol{H} - \boldsymbol{\mathcal{L}}\| \leq a + t_0^2 + 2t_0 \leq 3a. \qquad \blacksquare$$

REFERENCES

[1] E. Abbe, *Community detection and stochastic block models: Recent developments*, Journal of Machine Learning Research, 18 (2018), pp. 1–86, http://jmlr.org/papers/v18/16-480.html.

[2] M. Avella-Medina, F. Parise, M. T. Schaub, and S. Segarra, *Centrality Measures for Graphons: Accounting for uncertainty in networks*, IEEE Transaction on Network Science and Engineering, (2018), https://doi.org/10.1109/TNSE.2018.2884235.

[3] F. Bunea and L. Xiao, *On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA*, Bernoulli, 21 (2015), pp. 1200–1230.

[4] E. J. Candes, M. B. Wakin, and S. P. Boyd, *Enhancing sparsity by reweighted l1 minimization*, Journal of Fourier Analysis and Applications, 14 (2008), pp. 877–905.

[5] F. Chung and M. Radcliffe, *On the spectra of general random graphs*, Electronic Journal of Combinatorics, 18 (2011), p. 215.

[6] A. Dasgupta, J. E. Hopcroft, and F. Mcsherry, *Spectral analysis of random graphs with skewed degree distributions*, in Annual Symposium on Foundations of Computer Science, 2004, pp. 602–610.

[7] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation. III*, SIAM Journal on Numerical Analysis, 7 (1970), pp. 1–46.

[8] G. B. Giannakis, Y. Shen, and G. V. Karanikolas, *Topology identification and learning over graphs: Accounting for nonlinearities and dynamics*, Proceedings of the IEEE, 106 (2018), pp. 787–807.

[9] D. Hayden, Y. H. Chang, J. Goncalves, and C. J. Tomlin, *Sparse network identifiability via compressed sensing*, Automatica, 68 (2016), pp. 9 – 17, https://doi.org/http://dx.doi.org/10.1016/j.automatica.2016.01.008.

[10] D. Hayden, Y. Yuan, and J. Goncalves, *Network identifiability from intrinsic noise*, IEEE Transactions on Automatic Control, PP (2016), pp. 1–1, https://doi.org/10.1109/TAC.2016.2640219.

[11] T. Hoffmann, L. Peel, R. Lambiotte, and N. S. Jones, *Community detection in networks with unobserved edges*, arXiv preprint arXiv:1808.06079, (2018).
[12] C. Honey, O. Sporns, L. Cammoun, X. Gigandet, J.-P. Thiran, R. Meuli, and P. Hagmann, *Predicting human resting-state functional connectivity from structural connectivity*, Proceedings of the National Academy of Sciences, 106 (2009), pp. 2035–2040.
[13] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge university press, 1990.
[14] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Transactions on Automatic Control, 48 (2003), pp. 988–1001.
[15] A. Julius, M. Zavlanos, S. Boyd, and G. J. Pappas, *Genetic network identification using convex programming*, IET Systems Biology, 3 (2009), pp. 155–166.
[16] B. Karrer and M. E. Newman, *Stochastic blockmodels and community structure in networks*, Physical review E, 83 (2011), p. 016107.
[17] L. Ljung, *System identification*, in Signal Analysis and Prediction, Springer, 1998, pp. 163–173.
[18] L. Ljung, *Perspectives on system identification*, Annual Reviews in Control, 34 (2010), pp. 1–12.
[19] L. Lovász, *Large Networks and Graph Limits*, vol. 60, American Mathematical Soc., 2012.
[20] N. Masuda, M. A. Porter, and R. Lambiotte, *Random walks and diffusion on networks*, Physics Reports, (2017).
[21] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, *Connecting the dots: Identifying network structure via graph signal processing*, IEEE Signal Processing Magazine, 36 (2019), pp. 16–43.
[22] D. Materassi and M. V. Salapaka, *On the problem of reconstructing an unknown topology via locality properties of the Wiener filter*, IEEE Transactions on Automatic Control, 57 (2012), pp. 1765–1777.
[23] A. Mauroy and J. Hendrickx, *Spectral identification of networks using sparse measurements*, SIAM Journal on Applied Dynamical Systems, 16 (2017), pp. 479–513.
[24] R. Olfati-Saber, J. A. Fax, and R. M. Murray, *Consensus and cooperation in networked multi-agent systems*, Proceedings of the IEEE, 95 (2007), pp. 215–233.
[25] R. Overbeek, M. Fonstein, M. D'souza, G. D. Pusch, and N. Maltsev, *The use of gene clusters to infer functional coupling*, Proceedings of the National Academy of Sciences, 96 (1999), pp. 2896–2901.
[26] T. Qin and K. Rohe, *Regularized spectral clustering under the degree-corrected stochastic blockmodel*, in Advances in Neural Information Processing Systems, 2013, pp. 3120–3128.
[27] K. Rohe, S. Chatterjee, and B. Yu, *Spectral clustering and the high-dimensional stochastic blockmodel*, The Annals of Statistics, 39 (2011), pp. 1878–1915.
[28] G. Sanguinetti, J. Laidler, and N. D. Lawrence, *Automatic determination of the number of clusters using spectral algorithms*, in Workshop on Machine Learning for Signal Processing, 2005, pp. 55–60.
[29] M. T. Schaub, S. Segarra, and H. T. Wai, *Spectral partitioning of time-varying networks with unobserved edges*, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 4938–4942.
[30] S. Segarra, A. G. Marques, G. Mateos, and A. Ribeiro, *Network topology inference from spectral templates*, IEEE Transactions on Signal and Information Processing over Networks, 3 (2017), pp. 467–483.
[31] S. Segarra, A. G. Marques, and A. Ribeiro, *Optimal graph-filter design and applications to distributed linear network operators*, IEEE Transactions on Signal Processing, 65 (2017), pp. 4117–4131.
[32] S. Segarra, M. T. Schaub, and A. Jadbabaie, *Network inference from consensus dynamics*, in Decision and Control (CDC), 2017 IEEE 56th Annual Conference on, IEEE, 2017, pp. 3212–3217.
[33] R. Shafipour, S. Segarra, A. G. Marques, and G. Mateos, *Network topology inference from non-stationary graph signals*, in International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2017, pp. 5870–5874.
[34] S. Shahrampour and V. M. Preciado, *Reconstruction of directed networks from consensus dynamics*, in American Control Conference (ACC), 2013, IEEE, 2013, pp. 1685–1690.
[35] S. Shahrampour and V. M. Preciado, *Topology identification of directed dynamical networks via power spectral analysis*, IEEE Transactions on Automatic Control, 60 (2015), pp. 2260–2265.
[36] P. Spirtes and K. Zhang, *Causal discovery and inference: Concepts and recent methodological advances*, Applied Informatics, 3 (2016), p. 3.
[37] M. Timme and J. Casadiego, *Revealing networks from dynamics: an introduction*, Journal of Physics A: Mathematical and Theoretical, 47 (2014), p. 343001.

[38] J. N. Tsitsiklis, *Problems in decentralized decision making and computation.*, PhD thesis, Massachusetts Institute of Technology, Cambridge, 1984.
[39] U. Von Luxburg, *A tutorial on spectral clustering*, Statistics and Computing, 17 (2007), pp. 395–416.
[40] H.-T. Wai, A. Scaglione, and A. Leshem, *Active sensing of social networks*, IEEE Transactions on Signal and Information Processing over Networks, 2 (2016), pp. 406–419.
[41] H.-T. Wai, S. Segarra, A. E. Ozdaglar, A. Scaglione, and A. Jadbabaie, *Blind community detection from low-rank excitations of a graph filter*, arXiv preprint arXiv:1809.01485, (2018).
[42] H.-T. Wai, S. Segarra, A. E. Ozdaglar, A. Scaglione, and A. Jadbabaie, *Community detection from low-rank excitations of a graph filter*, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4044–4048.
[43] T. Xiang and S. Gong, *Spectral clustering with eigenvector selection*, Pattern Recognition, 41 (2008), pp. 1012–1029.
[44] Y. Yu, T. Wang, and R. J. Samworth, *A useful variant of the Davis-Kahan theorem for statisticians*, Biometrika, 102 (2014), pp. 315–323.
[45] Y. Yuan, G.-B. Stan, S. Warnick, and J. Goncalves, *Robust dynamical network structure reconstruction*, Automatica, 47 (2011), pp. 1230–1235.