Bernoulli **31**(1), 2025, 584–608
https://doi.org/10.3150/24-BEJ1740

Tractably modelling dependence in networks beyond exchangeability

WEICHI WU[1,a], SOFIA OLHEDE[2,b] and PATRICK WOLFE[3,c]

[1]*Center for Statistical Sciences,Department of Industrial Engineering, Tsinghua University, Beijing, China,* [a]*wuweichi@mail.tsinghua.edu.cn*
[2]*Institute of Mathematics, Chair of Statistical Data Science, EPFL, Lausanne, Switzerland,* [b]*sofia.olhede@epfl.ch*
[3]*Departments of Statistics and Computer Science, Purdue University, West Lafayette, USA,* [c]*patrick@purdue.edu*

We propose a general framework for modelling network data that is designed to describe aspects of non-exchangeability, with an explicit parameter describing the degree of non-exchangeability. Conditional on latent (unobserved) variables, the edges of the network are generated by their finite growth history (via a latent order) while the marginal probabilities of the adjacency matrix are modeled by a generalization of a graph limit function (or a graphon). In particular, we study the estimation, clustering and degree behavior of the network in this setting. We determine (i) the least squares estimator of a composite graphon attaining the minimax rate under weak dependence with respect to squared error loss; (ii) that spectral clustering is able to consistently detect the latent membership when the block-wise constant composite graphon is considered under additional conditions; and (iii) we are able to construct models with heavy-tailed empirical degrees under specific scenarios and parameter choices. We find conditions under which the spectral clustering is consistent under non-exchangeability, revealing that the application scope of classification can be broader than classic $i.i.d.$ or exchangeable assumptions. In aggregate, we explore why and under which general conditions non-exchangeable network data can be described by a stochastic block model. The new modelling framework is able to capture empirically important characteristics of network data such as sparsity combined with heavy tailed degree distribution, and add understanding as to what generative mechanisms will make them arise.

Keywords: Exchangeable arrays; nonlinear stochastic processes; statistical network analysis; stochastic block model

1. Introduction

Despite a concerted effort by network researchers to propose new flexible network models, the field remains challenged by a lack of realism in many of the generating mechanisms it employs. There is a limited understanding of the consequences of this misfit between models and data. Features common in real-world networks are not likely to occur generated by popular network models, a much remarked upon observation. This has yielded considerable efforts to generate permutation invariant arrays with an excess of stars and other sub-graphs to generate additional network heterogeneity.

The realisation that any infinite dense array with a permutation invariant distribution can be approximated by a stochastic blockmodel [8,37] has explained the popularity of the latter model in practice, but we do not know much of the properties of such fitted descriptions outside the exchangeable setting. We note here also the results of [50], made in a similar vein. The blockmodel is of course normally misspecified, as it is unrealistic to assume that real-world networks are exchangeable arrays, but in addition to approximating a function by block constants, we are also ignoring potentially remaining effects corresponding to residual correlation, and other sampling effects that we shall now explore.

A number of consequences follow from the assumption of exchangeability that fully determines the sampling properties of a generated array. In real world scenarios these sampling properties can be

1350-7265 © 2025 ISI/BS

Table 1. Estimated connection probability and p-values of Ljung-Box tests for blocks of the Facebook Data

Blocks	1	2	3	4	5	6	
p-value	$< 10^{-4}$	0.002	0.835	$< 10^{-4}$	0.088	0.453	
connect prob	0.746	0.098	0.011	0.116	0.146	0.037	
Blocks	7	8	9	10	11	12	13
p-value	$< 10^{-4}$	$< 10^{-4}$	0.807	0.001	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
connect prob	0.074	0.389	0.109	0.196	0.712	0.13	0.702

violated for a number of reasons. This motivates us to seek to understand what generating mechanisms are required to produce the data features that we see in practice. One consequence of exchangability is that if the data can be described by a blockmodel, then conditional on block membership, there should be no or limited residual correlation between edges.

To study this effect, a Facebook network data with 362 nodes (see [34]) is analysed via the non-parametric tool of the network histogram [37]. The network histogram consistently estimates the conditional expectation of the adjacency matrix (an enumeration of edge variables) given the latent variables using block constant probability surfaces. We then vectorize the sub-matrix according to an order relying on the degrees of nodes in the same way as the usual preferential attachment model. More precisely, the nodes with larger degrees enters the network earlier. Upon arrival, each node realizes its connection with the earlier nodes. Following the order of the realization of connections, we obtain an ordered vector of 0 and 1. Notice that the order we considered is a popular proxy of realization time of the edge variables in network studies, for example in the preferential attachment models. We then perform the Ljung-Box test in each block to determine the independence of the elements in the vector and present our results in the following Table 1. The length of the vector corresponds to the first 12 groups is 378 and the size of group 13 is 325. We also estimate the connection probability in each block using the proportion of edge variables that equal to 1. From Table 1, the exchangeability is a misspecified model assumption at 0.05 significance level for the considered network, since if the exchangeable assumption exactly holds most of the p values will be larger than $0.05/13$ due to the Bonferroni correction. The question then naturally arises; if our network adjacency matrix does not appear to be a perfectly exchangeable array, can we trust the estimated stochastic block representation?

To capture additional heterogeneity research has therefore focused on relaxing models away from standard forms of exchangeability [13,14,50], often modelling edge variables instead of formulating models solely in terms of the network nodes. In this paper we propose a mechanism that can mimic temporal network growth, and is based on the popular graphon model [8,37]. We shall call our model the ‘composite graphon model’, a model class we introduce and describe in detail in Section 2. Our understanding is encapsulated by using models of latent dependence, and we explore the performance of standard network algorithms with data produced from such a generative model. These networks exhibit power-law degrees and significant data heterogeneity, typical observed features of non-exchangability. The common types of data that could require such models include for example citation networks [27], ecological networks [46], technological networks such as the powergrid [39] or communications networks [15].

Mimicking the mechanism of network growth, or network evolution, to produce an output network, is a very general idea. This idea can be said to be the genesis of other very popular frameworks such as the Barabasi–Albert network scheme [5], and temporal evolution underpins the graph processes of Borgs et al. [12]. Borgs et al. [12] clarify the complementary relationship of their models to those of [50], whose modelling framework is nearly identical, even if their achieved results are naturally complementary. Our aim is different: by introducing a model class that has a simple dependence parameter, the variable χ_n which quantifies the degree of exchangeability through controlling the probability of a run in consecutively generated edges, and then understanding the effects on estimation in this setting for networks with n nodes, as is clarified by our Theorem 3.1.

Furthermore, it is important to study the application of standard network algorithms to non-standard network data, as in practice we can hardly check if conditions of exchangeability are satisfied. Motivated by usage of composite likelihood in classical inference, we study the impact of χ_n on estimation, clustering and degree patterns. We conclude that when χ_n is small, the exchangability assumption is adequate even if the true model is non-exchangeable, while the assumption is inadequate exactly when χ_n approaches unity. Therefore the results of this paper connects to certain facts of time series analysis. Notice that the estimator of connection probability of a stochastic block model is obtained by averaging the the edge variables $a_{ij}^{\prime}s$ assuming conditional independence among these edge variables given latent groups. When the conditional dependence arises and gets stronger, the convergence rate of the averaging estimator deteriorates. This is similar to estimating the mean of a first order stationary time series by its average. Even if the information of the order of the time series is missing, the consistency of such estimator is still guaranteed but the convergence rate worsens with stronger dependence. On the other hand, there is substantial difference between the nonexchangeable network and time series data. By nature the network data is double indexed so the certain order is able to produce heterogeneity under strong dependence, thus the effect of missing order for network can be more complicated than that for time series.

To be more concrete and granular, the contributions of this paper are fourfold. First, we obtain an estimator of the composite graphon model (as well as the composite version of the stochastic block model) attaining the minimax rate under weak dependence with respect to the $\mathcal{L}_2$ loss function. Nonparametric regression with stationary/non-stationary time series has already attracted increasing research attention; for example, see [25,55,62]. Our result can be considered as a network counterpart of [22].

Second, we investigate spectral clustering for the composite stochastic block model, which is the non-exchangeable counterpart of the existing results such as [43]. We find that the spectral clustering algorithm is robust to certain dependence structure of edges, which answers the question why the algorithm works well when the assumption of conditional independence for the Stochastic Block Model (SBM) fails, for example see [44]. In addition to in network science, spectral clustering has been applied in many scientific fields including image analysis, data mining and speech recognition, see for instance [26] and [6], where we do not know of any latent dependence strength.

Third, we construct a model with a heavy-tailed degree distribution by considering an unobserved latent order $\omega(\{\cdot,\cdot\})$ and certain dependence structure of the edge variables. This shows a new mechanism resulted from the missing information of correlation between edges that can produce a power law degree distribution that is not preferential attachment model [5,40] nor is it the inhomogeneous edge connection probability model [36], nor relying on the exchangeble point process on $\mathbb{R}_+^2$ [50,51].

Fourth, we establish a theoretical framework for the analysis of network data with latent time–order non–anticipatory edge structure by developing a concept closely related to the notion of dependence measure ([56]) in the literature of time series analysis in which area a similar framework has been successfully set up to accommodate non-stationarity (see for example [61]). This motivates us to develop many useful mathematical tools in this paper.

The paper is organized as follows. The composite graphon is introduced and investigated in Section 2. The least squares estimator for the composite graphon is analyzed in Section 3. The spectral clustering algorithm for the composite SBM is studied in Section 4. Section 5 contains the discussion and future work. Finally, the simulation results for spectral clustering algorithm and degree distribution, as well as the proofs of all results are relegated to the online supplementary material.

We first introduce the notations. For any set A, let $|A|$ denote the cardinality of A. For a positive integer n, we write $[n] = \{1,2,...,n\}$. For n–dimensional random vectors $\mathbf{v} = (v_i, 1 \le i \le n)$ and $\mathbf{u} = (u_i, 1 \le i \le n)$, write $(v_i, 1 \le i \le n) \stackrel{d}{=} (u_i, 1 \le i \le n)$ if $\mathbf{v}$ and $\mathbf{u}$ have the same distribution. For two numbers i, j, denote by $\{i, j\}$ the collection of i and j, e.g. $\{i, j\} = \{j, i\}$ and $\{i, i\} = \{i\}$. Whenever the notation $\{i, j\}$ appears, by default we assume that $i \ne j$. Let $[n] \times^* [n]$ denote the set $\{\{i, j\}, 1 \le i \le n, 1 \le j \le n, i \ne j\}$. Write as $\mathbf{1}(\cdot)$ the usual indicator function which is one if the corresponding event is true and zero otherwise. Write $a \wedge b$ for $\min(a,b)$, and $a \vee b$ for $\max(a,b)$. For a graph with adjacency matrix $A_{i,j}$, its marginal probability is the collection of $\{\mathbb{P}(A_{i,j} = 1), \{i, j\} \in [n] \times^* [n]\}$. Each $A_{i,j}$ indicates the presence of an edge between node i and j, and we refer to it as an edge variable. It is only an edge if $A_{i,j} = 1$. The joint probability of the graph adjacency matrix is $\mathbb{P}(A_{i,j} = a_{i,j}, \{i, j\} \in [n] \times^* [n])$ for $\{a_{i,j}, \{i, j\} \in [n] \times^* [n]\} \in \{0,1\}^N$ with $N = n(n-1)/2$. Let $0^0 = 1$ as per usual. For any two matrices $A = (a_{ij})_{i\in[n], j\in[n]}$ and $B = (b_{ij})_{i\in[n], j\in[n]}$, then $\|A\| = (\sum_{i\in[n], j\in[n]} a_{ij}^2)^{1/2}$ representing the $\mathcal{L}_2$ (Frobenius) norm, and $< A, B >= (\sum_{i\in[n], j\in[n]} a_{ij} b_{ij})^{1/2}$. For any vector $\mathbf{v} = (v_1, ..., v_d) \in \mathbb{R}^d$, let $\|\mathbf{v}\|_{\mathcal{L}_p} = \mathbb{E}(|\mathbf{v}|^p)^{1/p}$.

In the next section we will describe the technical framework, which is the composite graphon model, to quantify a network's departure from exchangeability. We note that a lack of exchangeability is a "non-property". "Non-properties" are notoriously hard to quantify. A network can be non-exchangeable in more than one possible manner. Our method of quantifying departure from exchangeability is merely a possible choice; a non-unique and possibly an imperfect choice. It allows us to quantify that under mild forms of non-exchangeability when standard network analysis tools are still applicable and useful.

2. The composite graphon model

We take inspiration from the non-linear Wold representation used in [56] to build graphs that have latent dependence structure.

Definition 2.1 (Composite graphon model). An undirected graph is a composite graphon model with respect to $i.i.d.$ latent variables $\{\xi_i\}_{1\le i\le n}$ if

(a) Conditioned on the latent variable $\xi = \{\xi_i\}_{1\le i\le n}$, there exists a finite $k \ge 1$, such that for all i, the conditional distributions of the edge variables $B_{\omega(\{i,j\})} = A_{i,j}, i \ne j$

$$\mathbb{P}(B_i | B_{i-1}, ..., B_{-\infty}, \xi) = \mathbb{P}(B_i | B_{i-1}, ..., B_{i-k}, \xi), \tag{1}$$

where $\omega(\{\cdot,\cdot\}) : [n] \times^* [n] \to [N]$ is a bijective map which represents the latent order of $A_{i,j}$, and edge variables $\{B_j, j \le 0\}$ correspond to the burn-in process, which could be chosen as $i.i.d.$ $Bernoulli(1/2)$ independent of $B_s, 1 \le s \le N$. Let l be the smallest k such that (1) holds. We also use the convention that for $k = 0$, $(B_{i-1}, ..., B_{i-k}) = \varnothing$ which corresponds to the independence.

(b) The marginal connection probabilities given ξ_i satisfy that

$$\theta_{i,j} := \mathbb{P}(A_{i,j} = 1 | \xi_1, ..., \xi_n) = f_n(\xi_i, \xi_j),\ i \ne j \tag{2}$$

for some symmetric function $f_n(\cdot,\cdot) \in [0,1]$.

We call $f_n(\cdot,\cdot)$ the composite graphon. If in addition $f_n(\cdot,\cdot)$ is block-wise constant in the $\mathbb{R}^2$ plane, then we have a composite stochastic block model (composite SBM).

The "burn-in process" has little impact on the network, rather like the starting values for a time series autoregressive (AR) model. (1) means that conditioning on the latent variables $\{\xi_i\}_{1\le i\le n}$, $(B_s =$

$A_{\omega^{-1}(s)})_{1\le s\le n(n-1)/2}$ forms an order l Markov chain. Therefore we call l the "memory parameter". When $l=0$, the composite graphon reduces to the usual graphon model. In fact, the likelihood of composite graphon/SBM is the composite likelihood [49] of graphon/SBM, which motives the name of model. The exchangeability property is deteriorated by the correlation between edges given the latent random variables if $l>0$. In the literature of community detection, stochastic block models (which can be produced by setting $l=0$ and fixed latent variables or piecewise block constant graphon functions) with correlated edges has been recently investigated by [57], where they assume multiple networks are observed. As pointed out by [57], dependency among edges are ubiquitous in real-world network data such as friendship and brain connectivity thus "the conditional independency assumption typically does not hold in practice and, therefore, could lead to a misspecified model". We also refer to [57] for relevant and concrete data examples. Recently non-vertex exchangeable models have attracted considerable attention. The majority of non-vertex changeability models considered there have certain exchangeable properties, such as "exchangeable partition probability function", edge exchangeability, exchangeable interarrival times, see for example [9,17,41]. As a comparison, our model does not rest on any particular exchangeable forms.

Note that there are network models such that the linkage probability of every edge variables depends on *all* the edge variable generated before it. Such models cannot be allowed by (1) since we require l is bounded. An important example is the preferential attachment model [5], which we will further discuss in Remark 2.2. Recently the asymptotic normality of the affine preferential attachment network models has been studied by [24].

The parameter of interest (2) is independent of the mapping $\omega(\{\cdot,\cdot\})$. This fact is crucial for estimating the composite graphon model without estimating $\omega(\{\cdot,\cdot\})$. Indeed, model (2) is quite flexible, including the usual graphon model as its special case. We then present a general pseudo algorithm for constructing the composite graphon model with memory parameter l. In the following, we omit the subscript n if this omission produces no potential for ambiguity. Notice that for $\omega(\{\cdot,\cdot\})$, $\omega^{-1}:[N]\to[n]\times^*[n]$ has the form $\omega^{-1}(k)=\{\{i,j\},\omega(\{i,j\})=k\}$. Write $U_{i-1}=(B_{i-1},\dots,B_{i-l})^T$, and for any series $\xi_i, i\in\mathbb{Z}$, denote by $\xi_{\{i,j\}}=\{\xi_i,\xi_j\}$ for short. From the generative pseudo–algorithm Algorithm 1, we see that the joint distribution of the edge variables of the composite graphon modelis fully determined by the following (infinite dimensional) parameters:

(i) $f(\omega^{-1}(i)), 1\le i\le N$,
(ii) $\mathbb{P}(B_i=1|U_{i-1}=u_{i-1},\xi)$ for $2\le i\le N$, $u_i\in\{0,1\}^l$ with constraints (3) for $2\le i\le N$.

In the classic case, the first of these specifications is solely via the graphon function, while specification in (ii) breaks the model exchangeability, and so makes the model more flexible. We recover the classic graphon model when $\mathbb{P}(B_i=1|U_{i-1}=u_{i-1},\xi)=\mathbb{P}(B_i=1|\xi)=f(\zeta_{\omega^{-1}(i)})$. For any composite graphon model (2), we define its associated composite graph as follows. Our model connects to the graph limit and convergence in the language of [33] in a marginal way similar to the "composite" concept in the classic statistics literature, see for instance [49]. Furthermore, [13] introduces a latent birth time concept, which is similar to our latent order concept in that it is also temporal. The differences lie in the fact that their latent birth time is for each vertex, while our latent order is for each edge variable, and more fundamentally, lie in the procedure that they drop this latent birth time in their final step to make the model exchangeable such that the labels carry no information while our model is not exchangeable by assuming the information of the labels is missing. As a consequence, [13] involves additional cost for the exchangeability.

We now further discuss the dependence resulted by (1). Consider any graph satisfying (1) with memory parameter l and w.r.t. the map $\omega(\{\cdot,\cdot\})$. Denote by $\mathcal{X}\subset\mathbb{R}^l$ the set of l-dimensional binary vectors with all entries 0 or 1. The following proposition shows that the dependence between $A_{i,j}$ and $A_{k,l}$ decreases as $|\omega(\{i,j\})-\omega(\{k,l\})|$ increases.

Algorithm 1 Generative Pseudo Algorithm

Generate $\xi_1,\ldots,\xi_n$. Calculate $f(\xi_i,\xi_j)$. Generate $i.i.d.\ Bernoulli(1/2)\ B_{-l},...,B_0$ what are independent of $\xi_i' s$ and $B_j' s, j \geq 1$

For given ω, generate B_1 as $\mathbb{P}(B_1 = 1|\xi) = f(\xi_{\omega^{-1}(1)})$.

Generate B_2 by two parameters $\mathbb{P}(B_2 = 1|B_1 = 1,\xi)$ and $\mathbb{P}(B_2 = 1|B_1 = 0,\xi)$, which satisfies that $f(\xi_{\omega^{-1}(2)}) = \mathbb{P}(B_2 = 1, B_1 = 1|\xi) + \mathbb{P}(B_2 = 1, B_1 = 0|\xi)$, where

$$\mathbb{P}(B_2 = 1, B_1 = 1|\xi) = f(\xi_{\omega^{-1}(1)})\mathbb{P}(B_2 = 1|B_1 = 1,\xi),$$

$$\mathbb{P}(B_2 = 1, B_1 = 0|\xi) = (1 - f(\xi_{\omega^{-1}(1)}))\mathbb{P}(B_2 = 1|B_1 = 0,\xi).$$

By using $\mathbb{P}(B_2 = 0|B_1,\xi) = 1 - \mathbb{P}(B_2 = 1|B_1,\xi)$, in step 2 we have constructed a two-dimensional multivariate Bernoulli model (B_1, B_2).

for $i = 1$ to $N - 1$ **do**

Generate B_{i+1} by parameters $\mathbb{P}(B_{i+1} = 1|U_i = u_i,\xi)$ which satisfy the following constraints:

$$\mathbb{P}(B_{i+1} = 1|\xi) = f(z_{\omega^{-1}(i+1)}) = \sum_{u_i} \mathbb{P}(B_{i+1} = 1|U_i = u_i,\xi)\mathbb{P}(U_i = u_i|\xi), \tag{3}$$

for $u_i \in \{0,1\}^l$, where $\mathbb{P}(U_i = u_i|\xi)$ could be obtained by the l dimensional multivariate Bernoulli model generated in previous iteration.

end for with output $B_1, B_2, \ldots, B_N$.

Proposition 2.1. *Consider any composite graphon model with memory parameter l and mapping $\omega(\{\cdot,\cdot\})$. Assume $\min_{a,b\in\mathcal{X}\times\mathcal{X}} \mathbb{P}(U_i = a|U_{i-l} = b,\xi) \geq \alpha' > 0$. Then uniformly for i,k and for all $u_i, u_{i-k} \in \mathcal{X}$, we have that for $k \geq l$,*

$$\mathbb{P}(U_i = u_i|U_{i-k} = u_{i-k},\xi) - \mathbb{P}(U_i = u_i|\xi) = O(\chi_1^k \tilde{p}),$$

where $\chi_1 = (1 - 2\alpha')^{1/l}$, the latent variable ξ is defined in Definition 2.1, and

$$\tilde{p} = \max_{1\leq s\leq l} \max_{1\leq i\leq n} \max_{a} |\max_{b} \mathbb{P}(U_i = a|U_{i-s} = b,\xi) - \min_{b} \mathbb{P}(U_i = a|U_{i-s} = b,\xi)|.$$

Notice that if given ξ, $l = 0$ and $\{B_i\}$ is an independent series, then $\tilde{p} = 0$ and the composite graphon reduces to the standard exchangeable graphon model.

When the data sets consist of a sequence of composite graphon, Proposition 2.1 also relates to the dependence strength of the sequence. Specifically, consider a sequence of graphs $\{G_n\}$ with adjacency matrices $A_n = \{A_{i,j,n}\}_{1\leq i,j\leq n}$. For a series of $1-1$ corresponding mappings ω_n: $[n] \times^* [n] \to [N]$, define the dependence measure of the adjacency matrices w.r.t $\omega_n(\{\cdot,\cdot\}) \in \mathbb{N}$ to be

$$\Delta_n(k) := \max_{s\in\{0,1\},i,j} |\mathbb{P}(B_{\omega_n(\{i,j\})} = s|\mathcal{F}_{\omega_n(\{i,j\})-k}) - \mathbb{P}(B_{\omega_n(\{i,j\})} = s)|, \tag{4}$$

where $\mathcal{F}_i = (B_{-\infty}, \ldots, B_i)$, and $B_{\omega_n(i,j)} = A_{i,j,n}$. We refer to smallest possible non-negative χ_n such that $\Delta_n(k) = O(\chi_n^k)$ as the *dependence parameter*. In the remaining of the paper, we write χ_n as χ if no confusion arises from this adoption. In Section 2.2 we study an example for which χ can be exactly calculated.

For convenience, we let $B_s, s \le 0$ follow an $i.i.d.$ Bernoulli(1/2) law and be independent of $\{B_s, 1 \le s \le N\}$. In our paper, we call $\{B_s\}_{1\le s\le N}$ the "ordered edge variables with respect to ω_n", while we call $A_{i,j,n}$ "edge variables". Note that for a sequence of graphs, their adjacency matrices form an array of dependent Bernoulli random variables, with the n_{th} row of the array corresponding to edge variables of a size n graph ordered by $\omega_n(\{\cdot,\cdot\})$. For each graph G_n, its edge variables behave as a time series indexed by $\omega_n(\{i,j\})$. The quantity produced by (4) is closely related to the physical and predictive dependence measure introduced by [56] that quantifies the degree of dependence of outputs on inputs in (nonlinear) physical systems. It is easy to compute for many stochastic process and has been used to quantify the strength of dependence in both stationary and non-stationary time series, see for example [55,62] among many others. We both introduce the dependence measure to networks, and use it to characterise dependence in our network sequence.

Graph sequence models have been well studied in the literature. Among others, for example, [10] studied the metrics for sparse graphs via a graph sequence model; [8] established a graph sequence model with a scaling parameter ρ_n to address the sparsity issues for exchangeable graph model; and [37] approximated the graphon model. The minimax rate of the estimation of sparse graphon sequence model was studied by [21,31] and others, see [23] for a detailed survey.

Remark 2.1 (Law of Graph Sequence). We provide two examples of graph sequences satisfying Komogorov's extension theorem. Thus we can define the law of $(G_n)_{n\ge 1}$. An obvious example is that G_{n+1} is independent of $(G_i)_{i\le n}$, corresponding to the observations from independent different network samples. In this case we have no restrictions on the network. Another example is the network growing sequence, in which case the latent order is compatible with the growth of network (the edge incident with node i is generated later than edge incident with nodes less than i, i.e. $w(i,j) > w(k,l)$, if $\max(i,j) > \max(k,l)$). Then we can obtain G_{n+1} from G_n via generating the new latent variable ξ_{n+1} independent of $\xi_1,\ldots,\xi_n$, and determine the conditional connecting probabilities via Algorithm 1. The conditions of Komogorov's extension theorem are satisfied due to the Markov structure of the edge generating mechanism.

By a straightforward argument using the Markov property, we have the following corollary:

Corollary 2.1. *Under the conditions of Proposition 2.1, we have that*

$$\Delta_n(k) = \max_{\substack{1\le i\le N,\\ b_s\in\{0,1\}}} |\mathbb{P}(B_i = b_i | B_s = b_s, s \le i-k, \xi) - \mathbb{P}(B_i = b_i|\xi)| = O(\chi_1^k), \tag{5}$$

where χ_1 is defined in Proposition 2.1.

By the definition of dependence parameter χ, we have $0 \le \chi \le \chi_1$. If $|\chi_1| \le 1-\epsilon$ for some $\epsilon > 0$, then we say the considered graph is short-range dependent w.r.t. $\omega(\{\cdot,\cdot\})$. In this case equation (5) implies a geometric decay of $|\mathbb{P}(B_i = b_i|B_s = b_s, s \le i-k, \xi) - \mathbb{P}(B_i = b_i|\xi)|$ in k. This shows a stronger link with an autoregressive process where the term $\mathbb{P}(B_i = b_i|B_s = b_s, s \le i-k, \xi)$ plays the role of conditional expectation of an observation on another observation k steps ahead from an autoregressive process.

2.1. Inhomogeneity of the composite graphon model

In this subsection, we explore the inhomogeneity introduced by conditional dependence via studying examples of composite SBM with memory parameter 1. The memory parameter is given by Definition 2.1. Note that the memory parameter 0 corresponds to the classical stochastic block model.

2.1.1. *Inhomogeneity introduced by communities*

We first consider the composite SBM with fixed k communities constructed as follows. Define the map $z : [n] \to [k]$, which assigns n nodes into k different groups. Define $\tilde{z} : [n] \times^* [n] \to [k] \times^* [k]$ as $\tilde{z}(i,j) = \{z(i), z(j)\}$. We shall construct a composite SBM such that the edge variable connection probability depends on its previous edge variable with respect to latent (and unobservable) map ω. For this purpose, let q, s be the numbers such that $\{q,s\} = \tilde{z}(\omega^{-1}(\omega(\{i,j\})-1))$, i.e., q,s are the communities of vertices of $B_{\omega(\{i,j\})-1}$. Assume given the value of $B_{\omega(i,j)-1}$ that

$$A_{i,j} = B_{w(\{i,j\})} \sim \begin{cases} \text{Bernoulli}(\varrho^{0,q,s}_{z(i),z(j)}) \text{ if } B_{\omega(i,j)-1} = 0 \\ \text{Bernoulli}(\varrho^{1,q,s}_{z(i),z(j)}) \text{ if } B_{\omega(i,j)-1} = 1 \end{cases}. \tag{6}$$

The probability of linking (i,j) depends on its "parent" edge variable $B_{\omega(i,j)-1}$, and the communities of the four nodes q, s, $\tilde{z}(i,j)$. Conditioning on the latent memberships, (6) reduces to an inhomogeneous two-state Markov process. Recently in time series analysis, researchers have developed certain inhomogeneous models to characterise non-stationarity of integer-valued and categorical data, see for example [47]. To define a composite SBM with k groups such that $f(\xi_i, \xi_j) = \theta_{z(i)z(j)}$ for $\frac{k(k+1)}{2}$ connection probabilities $\{\theta_{a,b}, a, b \in [k], \theta_{a,b} = \theta_{b,a}\}$, using Algorithm 1, we specify a composite SBM with the parameters $\{\varrho^{u,c,d}_{a,b}, u \in \{0,1\}, \{a,b,c,d\} \in [k]^4\}$ satisfying the following constraints:

(a) For $1 \le a \le b \le k$,

$$\theta_{a,b} = \frac{\varrho^{0,a,b}_{a,b}}{1 + \varrho^{0,a,b}_{a,b} - \varrho^{1,a,b}_{a,b}}.$$

(b) For $1 \le a \le b \le k$, $1 \le s \le q \le k$, $\{s,q\} \ne \{a,b\}$, $\varrho^{0,s,q}_{a,b}$ and $\varrho^{1,s,q}_{a,b}$ satisfy

$$\theta_{a,b} = \varrho^{0,s,q}_{a,b}(1 - \theta_{s,q}) + \varrho^{1,s,q}_{a,b}\theta_{s,q}.$$

In fact, each sub-chain with $(a,b) \to (a,b)$ describes a homogeneous Markov chain, i.e., if we consider any consecutively generated edge variables which connect the vertices that belong to the same pair of groups (a,b), then these edge variables form a homogeneous Markov Chain with stationary probability $(\varrho_{a,b}, 1 - \varrho_{a,b})$. If $k = 1$ (corresponding to the scenario of only one group), constraints (a), (b) degenerate to a strictly stationary 2-states Markov process. From this point of view, the inhomogeneity is introduced by the specification of communities with stationary probabilities $(\varrho_{a,b}, 1 - \varrho_{a,b})$.

2.1.2. *Inhomogeneity introduced by individuals*

Another source of inhomogeneity is due to the dependence introduced by the latent position $\omega(i,j)$. Consider the single group, or $k = 1$ case such that f is constant on $[0,1]^2$. As described by the algorithm under definition 2.1, another composite SBM could be specified by

$$\theta_{1,1} = \frac{p_{0,i}}{1 - p_{1,i} + p_{0,i}} = \frac{p_{0,j}}{1 - p_{1,j} + p_{0,j}} \quad \text{for } 1 \le i < j \le N, \tag{7}$$

where $p_{0,i} = \mathbb{P}(B_i = 1 | B_{i-1} = 0)$ and $p_{1,i} = \mathbb{P}(B_i = 1 | B_{i-1} = 1)$. It is not hard to see that (7) defines an inhomogeneous Markov chain. The conditional connection probabilities of the edge variables depend on their positions in the history of the Markov chain. Since all nodes belong to the same community, the inhomogeneity is evident only at the individual level. This is very different from its Erdős-Rényi counterpart, of which each node is stochastically equivalent.

2.2. Example: Marginally edge constant latent time–order graph

In this subsection, we study the effect of the posited conditional dependence by studying the given model of the marginally edge constant latent time-order graph sequence model (MECG). The MECG sequence model is defined as a composite graphon model with $f(\cdot,\cdot) \equiv c_n$ where c_n is a function of n, and $f(\cdot)$ is defined in (2). For each fixed n, $\mathbb{P}(A_{i,j} = 1) = \mathbb{P}(A_{k,l} = 1) = c_n$ for every $(i,j),(k,l), i \neq j, k \neq l$. It is also a composite SBM with only one group, (as the SBM with one group corresponds to the Erdös-Rényi model). Meanwhile, edge variables are correlated with respect to the latent order $\omega_n(\{\cdot,\cdot\})$. Via studying the MECG, we can investigate the effect of the (conditional) dependence separately from the effect of inhomogeneous (marginal) edge variables' linkage probabilities. In the following arguments, for simplicity we omit the subscript n of $\omega_n(\{\cdot,\cdot\})$, $p_{1,n}, p_{0,n}$ and p_n if no confusion arises.

Consider the Markov process of MECG, which we write as $CG(V,\omega,p_0,p_1)$:

$$\mathbb{P}(B_i|B_{i-1}) = \left(p_1^{B_i}(1-p_1)^{1-B_i}\right)^{B_{i-1}}\left(p_0^{B_i}(1-p_0)^{1-B_i}\right)^{1-B_{i-1}}, \tag{8}$$

$$\text{where } B_i = A_{\omega^{-1}(i)}.$$

Via (8), if $B_{i-1} = 1$ then it is distributed as $Bernoulli(p_1)$, otherwise it is distributed as $Bernoulli(p_0)$. By the fundamental theorem of Markov Chains, the edge variables of $CG(V,\omega,p_0,p_1)$ have a limiting distribution

$$p := \mathbb{P}(B_\infty = 1) = p_0/(1 + p_0 - p_1) \in [p_0 \wedge p_1, p_0 \vee p_1].$$

We then consider the stationary scenario, i.e.,

$$\mathbb{P}(B_j) = p \quad \forall 1 \le j \le N. \tag{9}$$

This is because when the total number of the edge variables is large, the majority of the edge variables of $CG(V,\omega,p_0,p_1)$ have marginal linkage probabilities close to p.

Definition 2.2. We say that the graph $CG(V,\omega,p_0,p_1)$ is a first order homogeneous MECG with vertices V, driven by the order $\omega(\{\cdot,\cdot\})$ if (8) and (9) hold.

Notice that when $p_0 = p_1 = p$, then the latent structure is not active. As a result, $CG(V,\omega,p_0,p_1)$ reduces to the standard Erdős-Rényi graph $G(|V|,p)$. The following corollary explicitly calculates the conditional probability of B_i given B_{i-k} for $k \ge 2$:

Corollary 2.2. *Consider $CG(V,\omega,p_0,p_1)$. Define $\mathbb{P}_k(a|b) = \mathbb{P}(B_j = a|B_{j-k} = b)$ for $a,b \in \{0,1\}$. Then we have that*

$$\mathbb{P}_k(1|0) = \frac{p_0(1-(p_1-p_0)^k)}{1-p_1+p_0}; \qquad \mathbb{P}_k(1|1) = \frac{p_0 + (1-p_1)(p_1-p_0)^k}{1-p_1+p_0},$$

$$\mathbb{P}_k(0|1) = \frac{(1-p_1)(1-(p_1-p_0)^k)}{1-p_1+p_0}; \qquad \mathbb{P}_k(0|0) = \frac{(1-p_1)+p_0(p_1-p_0)^k}{1-p_1+p_0}.$$

If we let $p = \frac{p_0}{1-p_1+p_0}$, then we have that

$$\mathbb{P}_k(1|0) = p - p(p_1-p_0)^k, \ \ \mathbb{P}_k(1|1) = p + (1-p)(p_1-p_0)^k$$

$$\mathbb{P}_k(0|1) = (1-p) - (1-p)(p_1-p_0)^k, \ \ \mathbb{P}_k(0|0) = (1-p) + p(p_1-p_0)^k.$$

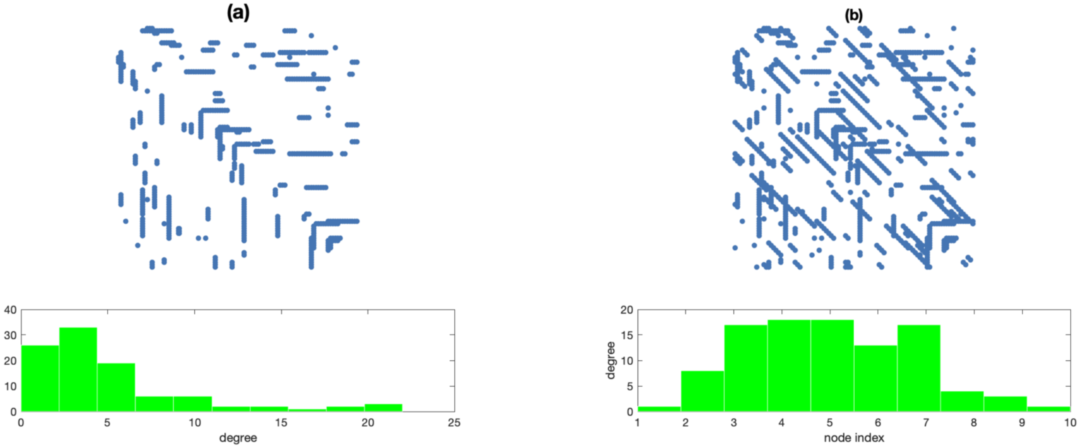

Figure 1. Adjacency matrix of $CG(V,\omega_1,p_0,p_1)$ (a) and $CG(V,\omega_2,p_0,p_1)$ (b), with a histogram of the degrees for the network below the adjacency spy plot. The histogram shows the "power-law" like behaviour of the (a) plot, relative to the (b) plot. The latent order ω_1 and ω_2 will be discussed in Sections 2.2.2 and 2.2.3, respectively.

The results show that the dependence between edge variables, or equivalently $\Delta_n(k)$ defined in (4) decays at the geometric rate $(p_1-p_0)^k$, i.e., $\chi_n = |p_1-p_0|$. We discuss the phase transition of MECG in Section C of the online supplementary material. In the remaining of the paper, we focus on the degree distribution of the MECG, to illustrate via degree sequence that when $\chi_n \longrightarrow 1$ the network model will have different degree patterns from the standard graphon models. Besides the degree distribution, there are other statistics behave differently in the composite and standard graphon models. For example, when $\chi_n \longrightarrow \chi$ for some $0<\chi<1$, consider the composite Erdos Renyi graph with $l \geq 2$. Then the expected number of triangles will not be p^3 as usual, where p is the edge connection probabilities, since $\mathbb{P}(A_{i,j}=1, A_{i,k}=1, A_{j,k}=1) \neq \mathbb{P}(A_{i,j}=1)\mathbb{P}(A_{i,k}=1)\mathbb{P}(A_{k,j}=1)$. The expected number of triangular depends on the parameter χ and latent orders. Due to the page limit, we shall leave the investigation of other statistics than degree distribution as a rewarding future work.

2.2.1. Degree distribution

In this subsection, by investigating simple examples, we shall see that i) different ordering $\omega_n(\{\cdot,\cdot\})$ have different impact on the network structure when $\chi_n \longrightarrow 1$, so the impact of missing information of $\omega_n(\{\cdot,\cdot\})$ is not asymptotic negligible when dependence is strong; and ii) our model is flexible enough to produce networks both with and without a heavy-tailed degree distribution. The idea is that when $\mathbb{P}(B_i=1|B_{i-1}=1) \longrightarrow 1$, we can design latent orders such that incident edge variables (e.g., $A_{i,j}$ and $A_{i,k}$) are strongly correlated (or weakly correlated), and hence their summation, or corresponding degrees, cannot (can) be well approximated by sums of independent Bernoulli random variables. To illustrate this, we display the adjacency matrix of typical $CG(V,\omega_1,p_0,p_1)$ and $CG(V,\omega_2,p_0,p_1)$ with network size $n=100, p_0=\frac{1}{n}, p_1=1-\frac{1}{n^{1/3}}$ in left and right panels of Figure 1. The latent order ω_1 and ω_2 will be discussed in Sections 2.2.2 and 2.2.3, respectively, and they produce very different degree patterns.

In the following, we shall construct two MECG which share the same $p_0=\lambda_0/n$ and p_1 such that $p_0 < p_1$ but different latent orders. In particular, the dependence are strong in the sense that the considered $\chi_n = p_1-p_0 \longrightarrow 1$. We shall show that one order leads to heavy-tailed degree distribution and the other order leads to light tail degree distribution, which indicates the complicated effect of latent orders when dependence is strong. It is also worth pointing out that the two examples are marginal sparse network in the sense that the marginal connection probability $p = \frac{p_0}{1-p_1+p_0}$ (by Corollary 2.2) approaches 0.

2.2.2. *Examples of MECG with heavy tail degree distribution*

Let $\varpi_1(i,j) = n(i-1) - i(i-1)/2 + j - i$ for $1 \le i < j \le n$, and $\omega_1(\{i,j\}) = \varpi(i \wedge j, i \vee j)$, $i \ne j$. Consider the first order homogeneous process MECG $CG(V,\omega_1,p_0,p_1)$. In particular we choose the ordering $\omega_1(\{i,j\})$ where the edge variables are generated as follows:

$$A_{1,2}, A_{1,3}, \ldots, A_{1,n}, A_{2,3}, \ldots, A_{2,n}, A_{3,4}, .., A_{3,n}, \ldots, A_{n-1,n}.$$

In other words, a node generates its edge variables after all the nodes labeling before it have generated their edge variables. We shall see that the considered ordering is able to generate a heavy-tailed degree distribution if we set $p_1 \to 1$, since a connected edge variables $A_{a,b}$ will lead to a high chance that the next edge variable $A_{a,b+1}$ is connected, where the two edge variables have the same vertex. In such a way our model is able to produce a larger number of high degree nodes than the Erdös-Rényi model.

We now study the empirical degree distribution $n^{-1}\sum_{i=1}^n I(d_i = k)$ for $1 \le k \le n$, where d_i is the degree of node i. When the nodes have a homogeneous degree distribution, for example in the Erdös-Rényi graph, $n^{-1}\sum_{i=1}^n I(d_i = k)$ is an unbiased estimator of $\mathbb{P}(d_1 = k)$, see for example [38]. Meanwhile, inhomogeneity introduced by strong dependence will distort the empirical degree distribution, i.e., as we shall show among a wide range of k, the expectation of $n^{-1}\sum_{i=1}^n I(d_i = k)$ of the graph decays with k at a polynomial rate. In this way, the graph displays the power law degree distribution. Different from the Erdös-Rényi model, $\mathbb{P}(d_i = k)$ for MECG is heterogeneous in i instead of remaining constant in i.

Theorem 2.3 (Heavy-tailed Degree Distribution). *Consider the first order homogeneous MECG $CG(V,\omega_1,p_0,p_1)$ with $|V| = n$. Suppose $p_0 = \frac{\lambda_0}{n}$ with $\lambda_0 \ge 1$, and $p_1 = 1 - \lambda_1 n^{-c}$, $0 < c < 1/2$. For any $\gamma > 1$, $\mu > 0$, define $M_\gamma : \sum_{k=1}^n M_\gamma \frac{1}{k^\gamma} = 1$, and $M_{\gamma,\mu} : \sum_{k=1}^n M_{\gamma,\mu}\frac{1}{k^\gamma}\exp(-\mu k) = 1$. Let $\mathcal{A}_{n,\gamma} = \{k : n^{-1}\sum_{i=1}^n \mathbb{P}(d_i = k) \ge M_\gamma k^{-\gamma}\}$, $\mathcal{B}_{n,\gamma,\mu} = \{k : n^{-1}\sum_{i=1}^n \mathbb{P}(d_i = k) \ge M_{\gamma,\mu} k^{-\gamma}\exp(-\mu k)\}$. Then there exist $a_0, b_0, c_0, d_0 > 0$ (which may depend on γ), such that*

$$\left\{k : \lfloor a_0 n^{\frac{2c}{1+\gamma}} \rfloor \le k \le \lfloor b_0 n^c \log n \rfloor\right\} \subset \mathcal{A}_{n,\gamma},$$

$$\{k : \lfloor c_0 \log n \rfloor \le k \le \lfloor d_0 n \rfloor\} \subset \mathcal{B}_{n,\gamma,\mu}.$$

Theorem 2.3 shows that, within a wide range of values of k, the tail of the distribution of the degrees of the MECG $CG(V,\omega_1,p_0,p_1)$ model behaves similarly to the power law distribution (or to a power law degree distribution with exponential cutoff, see [35]). Consider the usual Erdös-Rényi graph $G(V,p)$, where $p = \frac{p_0}{1-p_1+p_0}$, so that the marginal linkage probabilities of edge variables are the same as the first order homogeneous MECG $CG(V,\omega_1,p_0,p_1)$. Let $C_n = \{k : k \ge g(n)\}$ where $g(n) \to \infty$ arbitrarily slowly. By proposition D.2 (a Poisson approximation) in the supplementary material and the large deviation theorem (see the proof of Theorem 2.4 in the supplementary material), it follows that there exist constants c, d such that both $C_n \cap \mathcal{A}_{n,\gamma}$ and $C_n \cap \mathcal{B}_{n,\gamma,\mu}$ are subsets of $\{k : \lfloor an^c \rfloor \le k \le \lfloor bn^c \rfloor\}$ when n is large enough. Thus, the first order homogeneous MECG $CG(V,\omega_1,p_0,p_1)$ has much larger $|C_n \cap \mathcal{A}_{n,\gamma}|$ and $|C_n \cap \mathcal{B}_{n,\gamma,\mu}|$ than the simple random graph $G(V,p)$.

2.2.3. *Examples of MECG with light-tailed degree distribution*

In this section, we construct a first order homogeneous MECG $CG(V,\omega_2,p_0,p_1)$ which has similar $|C_n \cap \mathcal{A}_{n,\gamma}|$ and $|C_n \cap \mathcal{B}_{n,\gamma,\mu}|$ to that of Erdös-Rényi graph $G(V, p = \frac{p_0}{1-p_1+p_0})$. The order ω_2 we consider is $\omega_2(\{i,j\}) = \varpi_2(i \wedge j, i \vee j)$, $i \neq j$, where $\varpi_2(i,j) = i + \frac{(2n-(j-i))(j-i-1)}{2}$ for $1 \le i < j \le n$. In particular, our choices of ordering follow which the edge variables $A_{i,j}, i < j$ are generated are as follows

$$A_{1,2}, A_{2,3}, \dots, A_{n-1,n}, A_{1,3}, \dots, A_{n-2,n}, A_{1,4}, .., A_{n-3,n}, \dots, A_{1,n}.$$

Observe that the edge variables are generated in increasing order of $j-i$. Among edge variables with equal $j-i$, the edge variables with smaller i are generated earlier.

Theorem 2.4. *Consider the first order homogeneous MECG Graph $CG(V, \omega_2, p_0, p_1)$ where $p_0 = \frac{\lambda_0}{n}, \lambda_0 \ge 1$ and $p_1 = 1 - \lambda_1 n^{-c}, c \in (0, 1/2)$ are defined in Theorem 2.3. Let Y_n follow Poisson(np) for $p = \frac{p_0}{1-p_1+p_0}$. Let d_i be the degree of node i. Let $g(n)$ be a series of real numbers which diverges but may increase at an arbitrarily slow rate. The we have for some $\iota > c$, $\iota + c < 1$,*

$$(i) \sum_{k=0}^{\infty} |\mathbb{P}(d_i = k) - \mathbb{P}(Y_n = k)| = O(n^{\iota+c-1}),$$
$$(ii)\ \mathbb{P}(d_i = k) \le \exp(-0.5(\iota - c)k \log n) \text{ for } \lfloor n^{\iota} g(n) \rfloor \le k \le n. \tag{10}$$

Theorem 2.4 shows that the behavior of the MECG $CG(V, \omega_2, p_0, p_1)$ is similar to an Erdös-Rényi graph $G(n,p)$, in the sense that both of their degree distributions can be mimicked by a Poisson($\frac{\lambda_0}{\lambda_1} n^c$) random variable, for an appropriate specification of c. Equation (10) also shows that the tail of the degree distribution decays very rapidly. Theorems 2.3 and 2.4 shows the great flexibility and rich structure of our model class. We illustrate this in Figure 2. We discuss the images from left to right in Figure 2. Figure 2 shows typical graphs generated from MECGs $CG(V, \omega_1, p_0, p_1)$, $CG(V, \omega_2, p_0, p_1)$ and Erdős-Rényi $G(V,p)$, respectively with $p_0 = 0.01$, $p_1 = 1 - \frac{1}{n^{1/3}}$, $p = \frac{p_0}{1-p_1+p_0}$ with $|V| = n = 100$. From the figure, we see that the first network is very inhomogeneous: it has the most hubs among the three networks. The second network is less inhomogeneous than the first network, but is more inhomogeneous than the third network. Notice that we construct the three networks in such a way that the marginal connection probability is $n^{-2/3}$, where n is the size of the network. This is larger than the connectivity threshold $\frac{\log n}{n}$. However, the first network in Figure 2 has some isolated nodes just like the models of [13]. Since the marginal connection probability in our experiment is controlled, the expected total edges of the three networks are fixed. As a result, the structure with more hubs will also tend to have more small degree nodes, and also more isolated nodes.We refer to section B of the supplemental material for simulation studies on degree distributions of various MECGs.

Remark 2.2. A familiar model for networks with power law degree distributions is the preferential attachment (PA) model, where the network is growing sequentially node by node. In PA, a node can not affect the relationship among earlier nodes. This shares some features with our model. Thus, the generating order (or history) of the edge variables of PA could be written as

$$A_{1,2}, A_{1,3}, A_{2,3}, A_{1,4}, A_{2,4}, A_{3,4}, \dots, A_{1,n}, \dots, A_{n-1,n}.$$

with the associated ordering $\omega(\{i,j\}) = \varpi(i \wedge j, i \vee j)$, where $\varpi(i,j) = \frac{(j-1)(j-2)}{2} + i$, $1 \le i < j \le n$. The linkage probability of an edge variable is determined by the popularity of its earlier (more popular) vertices. Hence PA is not a latent time-order graph since the required properties fail to hold for any finite k by noticing that for PA model $\mathbb{P}(B_{w(\{i,j\})} | B_{w(\{i,j\})-1}, \dots, B_1)$ is determined by degrees of the node i (assuming that $i < j$ and j is the newcomer), which is a function of summations of a subset of $B_{w(\{i,j\})-1}, \dots, B_1$ that consists of edges incident on node i. The well-known heavy-tailed degree

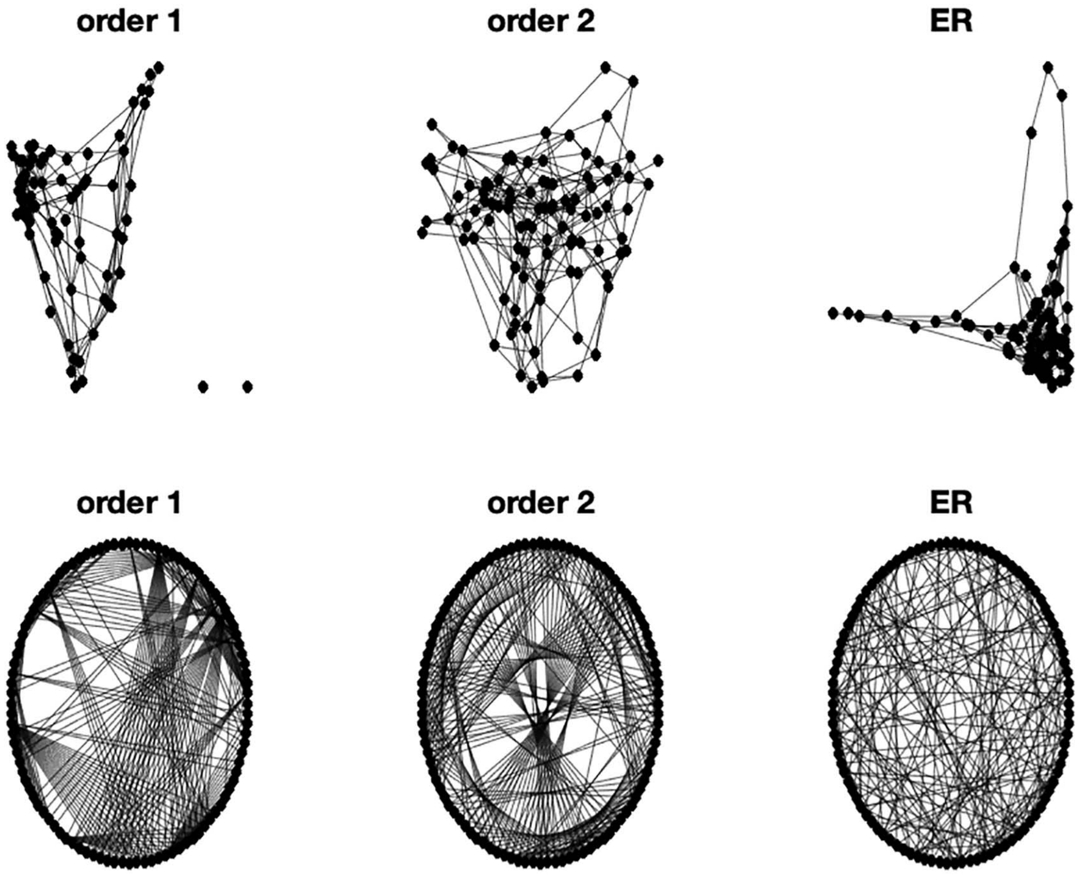

Figure 2. Left and Middle: typical visualization of the size $n = 100$ first order homogeneous MECG with ordering ω_1 and ω_2, and $p_0 = 0.01$, $p_1 = 1 - \frac{1}{n^{1/3}}$, respectively. Right: Size $n = 100$ Erdős-Rényi Graph with the $p = \frac{p_0}{1-p_1+p_0}$.

distribution of PA is contributed by infinite memory, order $\omega(\{,\})$, and inhomogeneous edge variables linkage probabilities.

Recently [14] proposed a class of normalized *unbounded* graphon model. Given latent positions $(\xi_i = x_i)_{1\le i\le n}$, the edge variables are independently connected with probabilities $\mathbb{P}(A_{ij} = 1|\xi_i = x_i, \xi_j = x_j) = \min(1, \rho W(x_i, x_j))$, where ρ is the target density and W is a (possibly) unbounded graphon. For detailed definition of ρ and W we refer to [11]. Due to the inhomogeneous conditional connection probabilities and the unboundedness of W, their model is allowed to have a large portion of high degree nodes and therefore the feature of heavy-tailed degree distribution under some circumstances.

In contrast to the aforementioned models, the MECG model $CG(V, \omega_1, p_1, p_0)$ has homogeneous (marginal) edge variable linkage probabilities. Hence the power law degree distribution is a consequence of the order $\omega_1(\{\cdot,\cdot\})$ and the strength of dependence which is determined by p_0, p_1. As a comparison, our construction $CG(V, \omega_2, p_1, p_0)$ does not have a power law distribution though it has the same strength of dependence as $CG(V, \omega_1, p_1, p_0)$. The only difference between the two MECGs is the order function. The unobserved ordering ω_i, $i = 1, 2$ introduces correlation, which increases the linkage probability of an edge variable between nodes adjacent in the ordering. This reveals the complex nature of the latent time-order graph, and has been illustrated by Figure 1. Notice that d_i is the sum of i_{th} row of the adjacency matrix. The figure shows that $CG(V,\omega_1,p_0,p_1)$ generates high degree nodes with greater frequency than $CG(V,\omega_2,p_0,p_1)$ due to the fact that degrees are calculated by summation along rows or columns, whilst diagonal structure does not aggregate to form larger degrees.

3. Least square estimator of the composite graphon

In this section, we assume that the memory parameter is bounded. Moreover, for $\alpha \in (0,1]$ and a sufficiently large constant M, we define the Hölder class

$$\mathcal{H}_\alpha(M) := \{f : |f(x,y) - f(x',y')| \le M(|x-x'| + |y-y'|)^\alpha, x \ge y, x' \ge y',$$
$$f \text{ symmetric}\},$$

for all $x \ge y$, $x' \ge y'$. Throughout the paper, we shall focus on the composite graphon model given in Definition 2.1 under either (A) or (B) of the following scenarios:

(A) $f(\cdot,\cdot)$ is a block-wise constant symmetric function, or
(B) $f \in \mathcal{F}_\alpha(M)$, where $\mathcal{F}_\alpha(M) = \{0 \le f \le 1 : f \in \mathcal{H}_\alpha(M)\}$.

Under (A), our model reduces to the composite stochastic block model (composite SBM). Under (B), the composite graphon $f(\cdot,\cdot)$ is smooth and estimable. The smoothness is assumed by for example [1,22,31,37] (which assumes $\alpha = 1$) among others. We discuss the least squares estimator of the composite graphon with respect to squared error loss. Let $\mathcal{Z}_{n,k} = \{z : [n] \to [k]\}$ be the collection of all possible mappings from $[n]$ to $[k]$. Then for any $\bar{z} \in \mathcal{Z}_{n,k}$, $\{\bar{z}^{-1}(a) : a \in [k]\}$ forms a partition of $[n]$, or equivalently: $\cup_{a\in[k]}\bar{z}^{-1}(a) = [n]$ and $\bar{z}^{-1}(a) \cap \bar{z}^{-1}(b) = \varnothing$ for any $a \ne b \in [k]$. In the following, we adopt the notation of [22]. Define for any matrix $(\eta_{ij})_{i\in[n],j\in[n]}$,

$$\bar{\eta}_{ab}(\bar{z}) = \frac{1}{|\bar{z}^{-1}(a)||\bar{z}^{-1}(b)|} \sum_{i\in\bar{z}^{-1}(a)} \sum_{j\in\bar{z}^{-1}(b)} \eta_{ij} \text{ for } a \ne b \in [k],$$

$$\bar{\eta}_{aa}(\bar{z}) = \frac{2}{|\bar{z}^{-1}(a)|(|\bar{z}^{-1}(a)|-1)} \sum_{i\in\bar{z}^{-1}(a), j\in\bar{z}^{-1}(a), i<j} \eta_{ij} \text{ for } a \in [k],\ |\bar{z}^{-1}(a)| > 1.$$

In this section, let $\theta_{i,j} = \mathbb{P}(A_{i,j} = 1|\xi) = f(\xi_i,\xi_j)$, where $A = A_{i,j}$ is the adjacency matrix. Define the estimate $\hat{\theta}_{ij} = \hat{Q}_{\hat{z}(i)\hat{z}(j)}$ where

$$(\hat{Q},\hat{z}) = \operatorname*{argmin}_{Q\in\mathbb{R}^{k\times k}_{sym}, \bar{z}\in\mathcal{Z}_{n,k}} L(Q,\bar{z}), \tag{11}$$

$$L(Q,\bar{z}) = \sum_{a,b\in[k]} \sum_{(i,j)\in\bar{z}^{-1}(a)\times\bar{z}^{-1}(b), i<j} (A_{ij} - Q_{ab})^2. \tag{12}$$

This procedure (11) is referred to as minimizing combinatorial least squares (see [22]). The word combinatorial is inserted, as we need to determine group membership (z) which is a combinatorial problem, rather than solely estimating a parameter by weighted averaging. Straightforward calculations show that $\hat{Q}_{ab} = \bar{A}_{ab}(\hat{z})$ for all $a,b \in [k]$. Therefore similarly to [53] we propose a block constant estimates for the composite graphon model. For this purpose, we define the space

$$\Theta_k = \{\{\theta^{\diamond}_{i,j}\} \in [0,1]^{n\times n} : \theta^{\diamond}_{i,i} = 0, \theta^{\diamond}_{i,j} = Q_{ab} = Q_{ba}, \forall (i,j) \in \bar{z}^{-1}(a) \times \bar{z}^{-1}(b), \text{ for some } Q_{ab} \in [0,1], \bar{z} \in \mathcal{Z}_{n,k}\}.$$

For each estimate $\hat{z}$, define $\tilde{Q}_{ab} \in [0,1]^{k\times k}$ by $\tilde{Q}_{ab} = \bar{\theta}_{ab}(\hat{z})$ and $\tilde{\theta}_{i,j} = \tilde{Q}_{\hat{z}(i)\hat{z}(j)}$ for $i \neq j$. For all $i \in \lfloor n \rfloor$, let $\hat{\theta}_{i,i} = \tilde{\theta}_{i,i} = \theta_{i,i} = 0$ (as we have assumed no self loops). Define $n_a = |\bar{z}^{-1}(a)|$. We first consider the composite SBM model. Define the true value on each block by $\{Q^*_{ab}\} \in [0,1]^{k\times k}$, and the oracle assignment $z^* \in \mathcal{Z}_{n,k}$, writing $\theta_{i,j} = Q^*_{z^*(i)z^*(j)}$.

Theorem 3.1. *Consider the composite SBM model G_n with k groups and dependence parameter χ. Assume the condition (A) hold. Then for any constant $C' > 0$, there is a constant $C > 0$ which only depends on C', such that*

$$\frac{1}{n^2} \sum_{i\in[n], j\in[n]} (\hat{\theta}_{i,j} - \theta_{i,j})^2 \leq C\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right)(1-\chi)^{-2}$$

with probability at least $1 - \exp(-C' n \log k)$ *uniformly over* $\theta \in \Theta_k$, *and*

$$\sup_{\theta\in\Theta_k} \mathbb{E}\left\{(\hat{\theta}_{i,j} - \theta_{i,j})^2\right\} \leq C_1\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right)(1-\chi)^{-2}$$

for all $k \in [n]$ with some universal constant $C_1 > 0$.

Note that when a graph sequence model is considered, the factor χ is allowed to depend on n. This is discussed further in Remark 3.1. Regarding the convergence rate, the term $\frac{k^2}{n^2}$ corresponds to the estimation of k^2 unknown parameters with an order of n^2 observations (edge variables), and the term $\frac{\log k}{n}$ corresponds to clustering rate, see for example [22] and [31]. Meanwhile, the term $(1-\chi)^{-2}$ is the effect of the non-exchangeability due to the latent order $\omega(\{\cdot,\cdot\})$ and the conditional dependence between edges given latent variables $\{\xi_i\}_{i\in\mathbb{Z}}$ w.r.t $\omega(\{\cdot,\cdot\})$. In this situation, the $(1-\chi)^{-2}$ part of the convergence rate degenerates to a constant, while the remaining part agrees with the rate in [22], which has been shown to be rate optimal.

Consider the composite graphon model (Algorithm 1) specified by $\{\xi_1,\ldots,\xi_n, A, \omega(\{\cdot,\cdot\}), l\}$ with symmetric composite grahon $f(\cdot,\cdot)$, i.e., $\theta_{i,j} = f(\xi_i,\xi_j)$, where latent variables $\{\xi_i, 1 \leq i \leq n\}$ are *i.i.d* $U(0,1)$, ω is the latent order of edge variables and l is the memory parameter. When $f \in \mathcal{F}_\alpha(M)$, which is the bounded Hölder's class defined in (B) of Section 2, arguments of Gao et al. (2015) show that there exists an *oracle* ([53]) $z^+_k \in \mathcal{Z}_{n,k}$ such that for some universal constant $C > 0$,

$$\frac{1}{n^2} \sum_{a,b\in[k]} \sum_{i\neq j: z^+_k(i)=a, z^+_k(j)=b} (\theta_{i,j} - \bar{\theta}_{a,b}(z^+_k))^2 \leq CM^2\left(\frac{1}{k^2}\right)^{\alpha\wedge 1}.$$

Define

$$(\theta^*, z^*) = \operatorname*{argmin}_{Q\in\mathbb{R}^{k\times k}_{sym}, z\in\mathcal{Z}_{n,k}} \tilde{L}(Q,z), \tilde{L}(Q,z) = \sum_{a,b\in[k]} \sum_{(i,j)\in\bar{z}^{-1}(a)\times\bar{z}^{-1}(b), i<j} (\theta_{i,j} - Q_{ab})^2.$$

By choosing $k = n^{\frac{1}{(1+\alpha\wedge 1)}}$, we have the following theorem.

Theorem 3.2. *Consider a composite graphon model $G = \{\xi_1,\ldots,\xi_n, A, \omega(\{\cdot,\cdot\}), l\}$ with dependence parameter χ. Assume the condition (B) hold. Then there exist constants C, C'*

$$\frac{1}{n^2}\sum_{i,j\in[n]}(\hat{\theta}_{i,j}-\theta_{i,j})^2 \le C(n^{\frac{-2(\alpha\wedge 1)}{1+\alpha\wedge 1}}+\frac{\log n}{n})(1-\chi)^{-2}$$

with probability at least $1-\exp(-C'n)$, uniformly over $f \in \mathcal{F}_\alpha(M)$. Furthermore,

$$\sup_{f\in\mathcal{F}_\alpha(M)} \mathbb{E}\left\{\frac{1}{n^2}\sum_{i,j\in\lfloor n\rfloor}(\hat{\theta}_{i,j}-\theta_{i,j})^2\right\} \le C_1(n^{\frac{-2(\alpha\wedge 1)}{1+\alpha\wedge 1}}+\frac{\log n}{n})(1-\chi)^{-2}, \tag{13}$$

for some constant $C_1 > 0$.

When χ_n is close to 1, the average of the adjacency matrix for a set of indices converges more slowly to its expectation, which impacts clustering, in which case the latent order matters. This is because if the edges within the same group are generated with latent orders close to each other, then their average could be quite different from their true means due to the strong dependence among them. On the other hand if the edges within the same group are generated with latent order far away from each other, then the dependence among them will be weaker than in the previous case so that their average could be close to their true means.

Remark 3.1. Comparing with the results of the usual SBM and the graphon model as discussed in [22], our Theorems 3.1 and 3.2 introduce an extra factor of $(1-\chi)^{-2}$ when the graph sequence model is considered. In particular, when $1-\chi \ge \eta > 0$, the rates of Theorems 3.1 and 3.2 achieve minimax rate as in [22]. The convergence rate is therefore slow when χ is close to 1, of which the situation indicates the strong conditional dependence between the edge variables given latent variables, see Proposition 2.1. A straightforward calculation using Proposition 2.1 shows that if given ξ_n, $l = 0$ and $\{B_i\}$ is an independent series, the rate is fully consistent with previous results in the sense that both the model and the rate recovers the known optimal rate of [22].

Remark 3.2. Consider the composite graphon model $G = \{\xi_1,\ldots,\xi_n, A, \omega(\{\cdot,\cdot\}), l\}$ with $l \ge 1$ investigated by Theorem 3.2. We construct a sparse composite graphon model

$$\tilde{G} = \{\xi_1,\ldots,\xi_n, \tilde{A}, \omega(\{\cdot,\cdot\}), l\}$$

by independently disconnecting the connected edge variables of the composite graphon model G with probability ρ_n. We therefore represent the "sparsity" by the parameter $0 < \rho_n \le 1$. This parameter was used by [10] to uniformly control the success probability across all nodes and uniformly controls the number of edges present. Straightforward calculations show that the upper bound of RHS of (13) could be lowered to the order of $\min_k\{\rho_n^2(\frac{1}{k^2})^{\alpha\wedge 1} + \rho_n(\frac{k^2}{n^2}+\frac{\log k}{n})(1-\chi)^2\}$, which coincides with the upper bound of that in [31]. In Section 2.2, we have displayed scenarios of homogeneity that for $i \ge 2$, $\frac{\mathbb{P}(B_i=1|B_{i-1}=1)}{\mathbb{P}(B_i=1|B_{i-1}=0)} \to \infty$ which is able to produce power law degree distribution. This scenario cannot be captured by scale parameter ρ_n. As a result, we do not focus on the scaled sparse model in detail.

Remark 3.3. Consider the graph sequence model. Assume that the conditions of Theorem 3.1 hold. By Theorem 3.1, when the number of communities k is fixed, a sufficient condition for the consistency of the $\mathcal{L}_2$ estimator (11) is $(\sqrt{n}(1-\chi))^{-1} = o(1)$. Meanwhile, Theorem 3.2 implies that when the composite graphon $f \in \mathcal{F}_\alpha(M)$ and $k = \lfloor n^{\frac{\alpha\wedge 1}{1+\alpha\wedge 1}} \rfloor$, a sufficient condition for the consistency of the $\mathcal{L}_2$ estimator (11) is $(n^{\frac{-2(\alpha\wedge 1)}{1+\alpha\wedge 1}} + \frac{\log n}{n})(1-\chi)^{-2} = o(1)$.

Interpretation: In other words, our estimator is consistent for $1-\chi = \Omega(n^{-a})$, $0 < a < 1/2$, where we write $a_n = \Omega(b_n)$ for series a_n, b_n if $b_n = O(a_n)$. By proposition 2.1 this means $\min_{a,b\in\mathcal{X}\times\mathcal{X}} \mathbb{P}(U_i = a|U_{i-l} = b, \xi) \geq Cn^{-a}$ for some constant C.

Remark 3.4. Theorem 3.2 determines that matrix estimation can be done for this problem, i.e. the sampled graphon can be estimated from an observed adjacency matrix. It does not necessarily relate to the underlying graphon function, unless we derive further results. As noted in [53] the mean square error of the estimate of $f(x,y)$ can be directly related to the matrix mean square error (see equation (E.3) in the online supplement). An issue with this statement is that the discretized $p_{ij} = f(\xi_i, \xi_j)$ is still random, as ξ_i is random and so statements can be made either marginally or conditionally on $\xi_1, \dots, \xi_n$. The added problem of estimating $f(x,y)$, the function, is to use an appropriate metric, and factor out measure preserving transformations.

It is discussed in [22] and [53] that the graphon model is closely related to non-parametric regression with unknown design and $i.i.d.$ errors. Consider the one-dimensional regression problem $y_i = f(z_i) + e_i$, where $z_i, 1 \leq i \leq n$ are $i.i.d.$ samples, and e_i are zero mean errors. When $f \in \mathcal{H}_\alpha(M)$ and e_i are $i.i.d$ normals, the local polynomial estimator achieves the minimax rate $n^{-\frac{2\alpha}{1+\alpha}}$ under the squared error loss $\frac{1}{n}\sum_{i\in[n]}\{\hat{f}(z_i) - f(z_i)\}^2$. When e_i is a short range dependent non-stationary time series for example the piecewise locally stationary time series in [60], Lemma 5 in [62] shows that $\mathbb{E}(e_i e_j) = O(\eta^{|i-j|})$ for some $\eta \in (0,1)$. It follows from this fact and Proposition 1.13 in [48], that the convergence rate of the local polynomial estimator with short-range dependent, non-stationary time series error has the same order as with $i.i.d.$ error.

However, under the situation that the design is unknown, an additional difference arises between the time series error and the $i.i.d$ error due to the unknown chronological order. Indeed, missing chronological order affects time series but not the $i.i.d.$ errors. Surprisingly, for the time series error, the impact of the missing chronological order on the estimation is negligible in terms of order under certain situation. Recall $L(Q,\bar{z})$ in (12), and define the new object $L^0(Q,\bar{z})$ by replacing A with $\theta = \mathbb{E}(A|\xi)$. For given $\bar{z}$, let $\tilde{Q}(\bar{z})$ and $Q^0(\bar{z})$ be the minimizer of $L(Q,\bar{z})$ and $L^0(Q,\bar{z})$, respectively. In fact, $\tilde{Q}(\bar{z})$ and $Q^0(\bar{z})$ are the average among partitions of adjacency matrix and of true but unknown conditional linkage probability matrix, respectively. Thus $\mathbb{E}(\tilde{Q}(\bar{z})) = Q^0(\bar{z})$. Since Bernoulli random variables are bounded, we show in our paper that under mild conditions and similar to the time series counterpart, the deviations between the average and the (conditional) mean of the edge variables are bounded uniformly over all possible partitions z under short-range dependence, as if the edge variables are conditional independent.

The time series structure of our model is important for many real applications. For example, [22] relates link prediction to the graphon model. In real application, links can be modeled by time series in dynamic network, see for example [45].

4. Spectral clustering algorithm for composite stochastic block model

In the previous sections we investigated the estimation of the composite graphon model and the composite stochastic block model. In addition to the estimation of the linkage probabilities, community detection is another research topic in network analysis. The connection between estimation and spectral clustering is complicated, and they are not identical problems. A good estimation result for the block heights of a stochastic blockmodel does not necessarily guarantee a good community detection result. For a more detailed discussion of the link between parameter estimation and spectral clustering, we refer to [22]. In the area of community detection, spectral clustering and its variants have already been widely applied [52]. The consistency of spectral clustering for certain exchangeable network models has been studied by for example [32,42,43,59] among others. In the following, we shall study the performance of spectral clustering for estimating the composite SBM.

4.1. Re-parameterization of the composite stochastic block model

By choosing a block-wise constant symmetric $f(\cdot,\cdot)$ in Definition 2.1, the composite SBM has the form of $\theta_{i,j} = \mathbb{P}(A_{i,j} = 1|\xi_i,\xi_j)$, where $\{\xi_i \in \mathbb{R}^k, 1 \le i \le n\}$ are $i.i.d.$ latent vectors, with one entry equal to one and all other entries equal to zero. Let θ be an $n\times n$ matrix, with $\theta_{i,j}$ its $(i,j)_{th}$ entry. Then θ could be parameterized as

$$\theta = ZB^\dagger Z^T,$$

where $B^\dagger \in [0,1]^{k\times k}$ is full rank and symmetric, and $Z \in \mathbb{R}^{n\times k}$ is a matrix with i_{th} row ξ_i such that it has one 1 in each row and at least one 1 in each column. For each node i, we say it belongs to group j if $\xi_{i,j}$, which is the j_{th} element of ξ_i, equals 1. With the re-parameterization, we are able to define the normalized graph Laplacian, which is essential for the spectral clustering algorithm. Define diagonal matrices D and $\bar{D}$ with diagonal elements $D_{i,i}$ and $\{\bar{D}_{i,i}\}, i = 1\ldots n$, respectively, where

$$D_{i,i} = \sum_{k=1}^{n} A_{i,k}, \quad \bar{D}_{i,i} = \sum_{k=1}^{n} \theta_{i,k}.$$

Define L and $\bar{L}$ for the Laplacian of A and θ, respectively, as

$$L = D^{-1/2}AD^{-1/2}, \bar{L} = \bar{D}^{-1/2}\theta\bar{D}^{-1/2}.$$

Note that $\bar{L}$ is the population version of L since the former is the Laplacian of θ and the latter is the Laplacian of adjacency matrix A. Both L and $\bar{L}$ may depend on the number of nodes n. Let $c_i = \bar{D}_{i,i}/n$ and $\tau_n = \min_{i=1,\ldots,n} c_i$. We shall write L as $L^{(n)}$, $\bar{L}$ as $\bar{L}^{(n)}$ and τ as τ_n when we need to emphasise the sample size. In the remainder of this section, we assume Z is unknown but fixed (unless specified). After obtaining L, the spectral clustering algorithm is given by:

1. Compute the eigenvectors $u_1,\ldots,u_k$ w.r.t. the first k largest eigenvalues of L in absolute value.
2. Run a k-means algorithm on vectors $y_1,\ldots,y_n$, $y_i \in \mathbb{R}^{1\times k}$ for $1 \le i \le n$ to cluster them into clusters $C_1,\ldots,C_k$, where y_i is the i_{th} row of matrix U, an $n\times k$ matrix such that the j_{th} column of U is u_k.

Then node i is in class g if y_i is assigned to C_g.

4.2. Properties of mis-clustered nodes

For simplicity of exploration and the ease of comparison, we will use the notion of [43]. In order to discuss the property of mis-clustered nodes, we first introduce the following notation. Let

$$P_n = \max_{j=1,\ldots,k} (Z^T Z)_{j,j},$$

which is the size of the largest cluster. We then give two properties of composite SBM. The validity of the properties could be shown similarly to Lemma 3.1 and Lemma 3.2 of [43], and so we omit the proof for the sake of brevity.

(a) There exists a matrix $V_1 \in \mathbb{R}^{k\times k}$ such that the columns of ZV_1 are the eigenvectors of $\bar{L}$ which correspond to the nonzero eigenvalues. In addition, $z_iV_1 = z_jV_1$ if and only if $z_i = z_j$.
(b) Let $V_2 \in \mathbb{R}^{n\times k}$ be a matrix whose orthonormal columns are the eigenvectors which correspond to the ordered largest k eigenvalues of L (in absolute value). Let $c_i, 1 \le i \le n$ be the centroid corresponding to the i_{th} row of V_2. Let the columns of $U, \bar{U} \in \mathbb{R}^{n\times k}$ be k orthonormal eigenvectors of LL and $\bar{L}\bar{L}$ (recall L and $\bar{L}$ are symmetric matrix) which correspond to the first k largest eigenvalues of the two matrices in absolute value, respectively. Define matrices O_1 and O_2 with the singular decomposition $\bar{U}^T U = O_1\Sigma O_2^T$, where O_1, O_2 are orthonormal matrices and Σ is a diagonal matrix. Let $O = O_1O_2^T$. Then $\|c_i - z_iV_1O\| < \frac{1}{\sqrt{2P_n}}$ if and only if $\|c_i - z_iV_1O\| < \|c_i - z_jV_1O\|$ for any $z_i \ne z_j$.

Under conditions of Theorem 4.1 below, the Davis-Kahan Theorem [20] shows that $\|V_2 - ZV_1O\| = o(1)$ almost surely, which leads to that the corresponding eigenvectors of the observed normalized graph Laplacian L is close to that of the population normalized graph Laplacian $\bar{L}$; see [43] for a detailed introduction of the Davis-Kahan Theorem. As a result, by (a), (b), we define the set of mis-clustered nodes as

$$\mathcal{M} = \left\{ i : \|c_i - z_iV_1O\| \ge \frac{1}{\sqrt{2P_n}} \right\},$$

since similarly to the argument in [43], we can show that if any node $i \notin \mathcal{M}$, then i will be correctly clustered by spectral clustering algorithm. Define $G(\chi, N) = \sum_{r=0}^{N} r^3\chi^{r/2}$, where χ is defined in Proposition 2.1.

Before stating Theorem 4.1 regarding the performance of spectral clustering for estimating the composite SBM, we present the following Proposition 4.1 which studies the tail probability of $\|LL - \bar{L}\bar{L}\|$. The latter is the difference between the population version of and the usual normalized graph Laplacian. The normalized graph Laplacian plays a central role in the spectral clustering, therefore the difference $\|LL - \bar{L}\bar{L}\|$ is key to study the asymptotic behavior of the corresponding clusters.

Proposition 4.1. *Consider a size n composite SBM with a fixed unknown mapping $\omega(\{\cdot,\cdot\})$ and dependence parameter χ. Denote by k_n the number of groups of nodes. Let $|\lambda_1| > \dots > |\lambda_{k_n}|$ be the absolute values of ordered k_n largest absolute and non-zero eigenvalues of $\bar{L}$. Assume that $n^{-1/2}(\log n)^2 G^{1/2}(\chi, N)(1-\chi)^{-1/2} = O(\lambda_{k_n}^2)$, and $\tau_n^2 > M/\log n$ for a sufficiently large constant M. Then there exist sufficiently large positive constants η_0, η_1, M' such that if $n \ge M'$*

$$\mathbb{P}(\|LL - \bar{L}\bar{L}\| \ge \frac{\log n}{\tau_n^2 n^{1/2}} G^{1/2}(\chi, N)(1-\chi)^{-1/2}) \le \zeta(n),$$

where

$$\zeta(n) = \frac{\eta_0}{n \log^4 n} + \eta_1 n^{-2}.$$

The proof of Proposition 4.1 is inspired by [43] and [61], see the online supplement for more details.

Theorem 4.1. *Under conditions of Proposition 4.1, we have that the number of miss-clustered nodes has the order of*

$$|\mathcal{M}| = o\left(\frac{P_n \log^2 n}{\lambda_{k_n}^4 \tau_n^4 n} G(\chi, N)(1-\chi)^{-1}\right), \quad a.s.$$

The conditions on the eigenvalues and on τ are similar to those of [43] that ensure the eigengap of $\bar{L}\bar{L}$ and the smallest nonzero eigenvalues of $\bar{L}$ cannot be too small. A referee provides an interpretation for the results of this theorem. Consider $u^T Av = \sum_{i,j} A_{ij} u_j v_i$ for some vectors u, v, then one can see that $u^T Av = \sum_{k=1}^N B_k x_k$ with $v_i u_j = x_k$ when $\omega_n(i,j) = k$. Then the process $(B_1 x_1, B_2 x_2, \dots)$ conditional on ζ is eventually nonstationary, but yet by some ergodic arguments it shall be that $u^T Av \approx \sum_k E(B_k|\zeta) x_k$.

Theorem 4.1 shows the consistency of the spectral clustering algorithm for composite SBM under regularity conditions. With χ gets closer to 1, the dependence between edge variables becomes stronger and the theoretically guaranteed convergence rate deteriorates. On the other hand, the requirement that $\tau_n^2 > M/\log n$ is almost as restrictive as the requirement of at least linearly growing expected degree for all nodes. Recently, many complex models have been proposed based on the SBM to capture additional and important graph structures. For instance, the general SBM proposed in [16] allows for a portion of arbitrary outliers, where the majority of nodes are generated from a fixed SBM. As a comparison, all nodes from the composite SBM in this paper differ from the SBM when edge variables are conditionally dependent given the latent membership. Another prominent model that can generate arbitrary degree inhomogeneity is the degree corrected stochastic block model (DC-SBM) (see [30]). For this model, consistency of community detection has been studied (see for example [59]), and corresponding spectral clustering algorithms have been proposed (for example see [42]). Also, SBM has been generalized to a mixed membership (for example [1]), and the K-median approach [58]. A tensor approach [4] have been proposed to estimate mixed-membership models. In this paper, we have built up a general framework for non-exchangeable graphs, and investigate the spectral clustering algorithm for composite SBM in detail. The extension of methods tailored to the corresponding composite generalized SBM are possible; for example one could consider SCORE [28] for composite degree-corrected SBM which can be obtained by involving more degree parameters based on composite SBM. Due to the page limit, we shall leave it as a rewarding future work.

Remark 4.1. The key concepts of a "composite graph" and composite SBM in Sections 2 and 4 are closely related to notion of composite likelihood. Composite likelihood inference is a popular and successful tool for statistical research when the joint likelihood is hard to evaluate, see [49] among others for a comprehensive review. In the literature of network analysis, the idea of analyzing pseudo or approximate likelihood has been proposed to tackle the complex and computational-infeasible joint likelihood of graph models, see for example ([2,3] and [7] among others.)

Remark 4.2. Assume that $k_n \equiv k$ and that there exists $c > 0$ such that $n_k \geq \lfloor cn \rfloor$, where n_k is the smallest size of group. Straightforward calculations show that $G(\chi, N)$ is of the order $(1-\chi^{1/2})^{-4}$. As a result, the condition $n^{-1/2}(\log n)^2 G^{1/2}(\chi, N)(1-\chi)^{-1/2} = O(\lambda_k^2)$ reduces to

$$(1-\chi)(1-\chi^{1/2})^{-4} \log^4 n = O(n), \tag{14}$$

which yields the weak consistency of clustering in the sense of [59], i.e, the mis-clustering rate in Theorem 4.1 is therefore simplified to $|\mathcal{M}| = o((1-\chi)(1-\chi^{1/2})^{-4} \log^4 n) = o(n)$. A straightforward calculation shows that a sufficient condition for (14) is that

$$1 - \chi = \Omega(n^{-1/5} \log n^{-4/5}).$$

As a comparison, Remark 3.3 shows that the estimation error of Theorem 3.1 is negligible if $1 - \chi = \Omega(n^{-1/2})$.

To close this section, we refer to Section A of the supplemental material for numerical studies for mis-clustering rate of spectral clustering algorithm for composite SBM model.

5. Discussion

As social media data sets, and other types of relational observations (networks) have become prevalent, so unsurprisingly the mathematical treatment of data taking the form of relationships between entities has become increasingly well-studied. The analysis of networks has been the focus of considerable efforts where the properties of estimators for popular models have now been established. Following on from the understanding of correctly specified parametric models is the usage of non-parametric and incorrectly specified models. For example, our understanding of classical approaches can be found to extend when considering dense exchangeable arrays, see for example [18,32,36,53].

Unfortunately the world contains many data sets that cannot be assumed to be exchangeable, despite how innocuous the assumption may seem, rather like that of stationarity for time series. For that reason we introduce the composite graphon model, and finite memory latent time-order graphs. By focusing on the latent variables in the model directly, we can build a continuum of types of networks that are exchangeable, or strongly non-exchangeable, all tuned explicitly in terms of the dependence strength. This helps us to understand data of this form, and when we can apply regular network tools to novel types of data, and understand the consequences of that choice.

Non-exchangeable networks produce many challenges. The presence of strong powerlaws in the degrees and further heterogeneity in the graphon function itself are still challenging researchers. It is not unreasonable to believe that these features reflect how the network was formed. By assuming that the network formed sequentially we are able to both define a parameter that tunes its degree of exchangeability, and thus we may understand standard tools when applied to such data. Our understanding of this mechanism simultaneously give glimpses into the formation of non-exchangeability, and provide a gray-scale understanding of networks, letting us see how the mechanism allows us to gradually "dial away" from exchangeability as a consequence of network evolution and growth.

A number of developments have sought to understand greater heterogeneity by modelling edge variables directly rather than relationships between nodes [19], this allowing a more natural and direct treatment of edge sparsity than some competing models. Others have concerned developing the practical application of work by Kallenberg's constructions [29], such as [12,13,17,50]. The two key aspects of the latter construction is to use a latent Poisson construction and a latent time. We also used a latent variable which is uniform rather than Poisson. We correlate the latent uniforms directly, and show how the correlation of the latent variables drive the degree of non-exchangeability directly and quantitatively. The advantage of our framework is that it naturally straddles the model space between strong heterogeneity to the standard exchangeable graph model, with a direct tuning of its degree of non-regularity. If the correlation is not too strong, then standard methods apply for estimating the graphon model, rather like in time-series analysis with short range dependent errors when estimating polynomial trends. As the correlation becomes very strong, the observations exhibit more strong heterogeneity, and standard tools like the stochastic blockmodel approximation of the underlying graphon model will become increasingly problematic.

We should note here also that our analysis is made for two different models; the composite SBM and the composite graphon. In the case of the former other methods could be applied to estimate the latent group structure, e.g. [32,58]. We conjecture that just like k–means, we could estimate a composite SBM using more sophisticated methods, and still arrive at consistent estimator.

A number of questions remain unanswered. In parts this falls back to the difficulty of understanding a non-property, which has already haunted both non-stationary and non-linear time series (there are many ways to be non-stationary or non-linear, but only one to be stationary). In parts it falls back to understanding non-exchangeability itself, as one property rather than several real-life observed consequences thereof. By providing this framework, we can better see the limitations of exchangeable models, and how exchangeability can fail to materialize as a consequence of dependence.

Acknowledgments

The authors would like to thank the editor, the associate editor and two anonymous reviewers for their helpful comments.

Funding

This work was supported by NSFC Program (No.12271287 and No.11901337) and the European Research Council under Grant CoG 2015-682172NETS, within the Seventh European Union Framework Program, and SCO and PJW acknowledge the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme Statistical Network Analysis where work on this paper was undertaken. This work was therefore also supported by EPSRC grant no. (EP/K032208/1).

Supplementary Material

Supplement to "Tractably modelling dependence in networks beyond exchangeability" (DOI: 10.3150/24-BEJ1740SUPP; .pdf). Supplementary material [54] provides simulation results for Sections 2.2 and 4. It also contains a discussion of the phase transition for connectivity and giant component of MECG, as well as the detailed proof of results in Sections 2-4.

References

[1] Airoldi, E.M., Costa, T.B. and Chan, S.H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Adv. Neural. Inf. Process. Syst.* 692–700.

[2] Amini, A.A., Chen, A., Bickel, P.J. and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41** 2097–2122. MR3127859 https://doi.org/10.1214/13-AOS1138

[3] Amini, A.A. and Levina, E. (2018). On semidefinite relaxations for the block model. *Ann. Statist.* **46** 149–179. MR3766949 https://doi.org/10.1214/17-AOS1545

[4] Anandkumar, A., Ge, R., Hsu, D. and Kakade, S.M. (2014). A tensor approach to learning mixed membership community models. *J. Mach. Learn. Res.* **15** 2239–2312. MR3231594

[5] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634 https://doi.org/10.1126/science.286.5439.509

[6] Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping Multidimensional Data* 25–71. Springer.

[7] Bickel, P., Choi, D., Chang, X. and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.* **41** 1922–1943. MR3127853 https://doi.org/10.1214/13-AOS1124

[8] 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.

[9] Bloem-Reddy, B. and Orbanz, P. (2017). Preferential attachment and vertex arrival times. ArXiv preprint. Available at arXiv:1710.02159.

[10] Bollobás, B. and Riordan, O. (2009). Metrics for sparse graphs. In *Surveys in Combinatorics 2009*. *London Mathematical Society Lecture Note Series* **365** 211–287. Cambridge: Cambridge Univ. Press. MR2588543

[11] Borgs, C., Chayes, J.T., Cohn, H. and Ganguly, S. (2021). Consistent nonparametric estimation for heavy-tailed sparse graphs. *Ann. Statist.* **49** 1904–1930. MR4319235 https://doi.org/10.1214/20-aos1985

[12] Borgs, C., Chayes, J.T., Cohn, H. and Holden, N. (2017). Sparse exchangeable graphs and their limits via graphon processes. *J. Mach. Learn. Res.* **18** Paper No. 210. MR3827098

[13] Borgs, C., Chayes, J.T., Cohn, H. and Veitch, V. (2019). Sampling perspectives on sparse exchangeable graphs. *Ann. Probab.* **47** 2754–2800. MR4021236 https://doi.org/10.1214/18-AOP1320

[14] Borgs, C., Chayes, J.T., Cohn, H. and Zhao, Y. (2019). An L^p theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions. *Trans. Amer. Math. Soc.* **372** 3019–3062. MR3988601 https://doi.org/10.1090/tran/7543

[15] Broido, A.D. and Clauset, A. (2019). Scale-free networks are rare. *Nat. Commun.* **10** 1017.

[16] Cai, T.T. and Li, X. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Ann. Statist.* **43** 1027–1059. MR3346696 https://doi.org/10.1214/14-AOS1290

[17] Caron, F. and Fox, E.B. (2017). Sparse graphs using exchangeable random measures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1295–1366. MR3731666 https://doi.org/10.1111/rssb.12233

[18] Choi, D.S., Wolfe, P.J. and Airoldi, E.M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284. MR2931253 https://doi.org/10.1093/biomet/asr053

[19] Crane, H. and Dempsey, W. (2018). Edge exchangeable models for interaction networks. *J. Amer. Statist. Assoc.* **113** 1311–1326. MR3862359 https://doi.org/10.1080/01621459.2017.1341413

[20] Davis, C. and Kahan, W.M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. MR0264450 https://doi.org/10.1137/0707001

[21] 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. MR3569248

[22] Gao, C., Lu, Y. and Zhou, H.H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43** 2624–2652. MR3405606 https://doi.org/10.1214/15-AOS1354

[23] Gao, C. and Ma, Z. (2021). Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing. *Statist. Sci.* **36** 16–33. MR4194201 https://doi.org/10.1214/19-STS736

[24] Gao, F. and van der Vaart, A. (2017). On the asymptotic normality of estimating the affine preferential attachment network models with random initial degrees. *Stochastic Process. Appl.* **127** 3754–3775. MR3707244 https://doi.org/10.1016/j.spa.2017.03.008

[25] Gao, J. (2007). *Nonlinear Time Series: Semiparametric and Nonparametric Methods*. *Monographs on Statistics and Applied Probability* **108**. Boca Raton, FL: CRC Press/CRC. MR2297190 https://doi.org/10.1201/9781420011210

[26] Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech Commun.* **16** 261–291.

[27] Ji, P. and Jin, J. (2016). Coauthorship and citation networks for statisticians. *Ann. Appl. Stat.* **10** 1779–1812. MR3592033 https://doi.org/10.1214/15-AOAS896

[28] Jin, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43** 57–89. MR3285600 https://doi.org/10.1214/14-AOS1265

[29] Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles*. *Probability and Its Applications (New York)*. New York: Springer. MR2161313

[30] Karrer, B. and Newman, M.E.J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E (3)* **83** 016107. MR2788206 https://doi.org/10.1103/PhysRevE.83.016107

[31] Klopp, O., Tsybakov, A.B. and Verzelen, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.* **45** 316–354. MR3611494 https://doi.org/10.1214/16-AOS1454

[32] Li, T., Lei, L., Bhattacharyya, S., Van den Berge, K., Sarkar, P., Bickel, P.J. and Levina, E. (2022). Hierarchical community detection by recursive partitioning. *J. Amer. Statist. Assoc.* **117** 951–968. MR4436325 https://doi.org/10.1080/01621459.2020.1833888

[33] Lovász, L. (2012). *Large Networks and Graph Limits*. American Mathematical Soc.

[34] Maier, B.F. and Brockmann, D. (2017). Cover time for random walks on arbitrary complex networks. *Phys. Rev. E (3)* **96** 042307. https://doi.org/10.1103/PhysRevE.96.042307

[35] Newman, M.E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46** 323–351.
[36] Olhede, S.C. and Wolfe, P.J. (2013). Degree-based network models. ArXiv preprint. Available at arXiv:1211.6537.
[37] Olhede, S.C. and Wolfe, P.J. (2014). Network histograms and universality of blockmodel approximation. *Proc. Natl. Acad. Sci. USA* **111** 14722–14727. https://doi.org/10.1073/pnas.1400374111
[38] Ouadah, S., Robin, S. and Latouche, P. (2020). Degree-based goodness-of-fit tests for heterogeneous random graph models: Independent and exchangeable cases. *Scand. J. Stat.* **47** 156–181. MR4075233 https://doi.org/10.1111/sjos.12410
[39] Pagani, G.A. and Aiello, M. (2013). The power grid as a complex network: A survey. *Phys. A* **392** 2688–2700. MR3043171 https://doi.org/10.1016/j.physa.2013.01.023
[40] Papadopoulos, F., Kitsak, M., Serrano, M.Á., Boguná, M. and Krioukov, D. (2012). Popularity versus similarity in growing networks. *Nature* **489** 537.
[41] Pitman, J. (2006). *Combinatorial Stochastic Processes*. *Lecture Notes in Math.* **1875**. Berlin: Springer. MR2245368
[42] Qin, T. and Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Adv. Neural. Inf. Process. Syst.* 3120–3128.
[43] Rohe, K., Chatterjee, S. and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887
[44] Saldaña, D.F., Yu, Y. and Feng, Y. (2017). How many communities are there? *J. Comput. Graph. Statist.* **26** 171–181. MR3610418 https://doi.org/10.1080/10618600.2015.1096790
[45] Sarkar, P., Chakrabarti, D. and Jordan, M. (2014). Nonparametric link prediction in large scale dynamic networks. *Electron. J. Stat.* **8** 2022–2065. MR3273618 https://doi.org/10.1214/14-EJS943
[46] Solé, R.V. and Montoya, J.M. (2001). Complexity and fragility in ecological networks. *Proc. Biol. Sci.* **268** 2039–2045. https://doi.org/10.1098/rspb.2001.1767
[47] Truquet, L. (2019). Local stationarity and time-inhomogeneous Markov chains. *Ann. Statist.* **47** 2023–2050. MR3953443 https://doi.org/10.1214/18-AOS1739
[48] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. MR2724359 https://doi.org/10.1007/b13794
[49] Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42. MR2796852
[50] Veitch, V. and Roy, D.M. (2015). The class of random graphs arising from exchangeable random measures. ArXiv preprint. Available at arXiv:1512.03099.
[51] Veitch, V. and Roy, D.M. (2019). Sampling and estimation for (sparse) exchangeable graphs. *Ann. Statist.* **47** 3274–3299. MR4025742 https://doi.org/10.1214/18-AOS1778
[52] von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. MR2409803 https://doi.org/10.1007/s11222-007-9033-z
[53] Wolfe, P.J. and Olhede, S.C. (2013). Nonparametric graphon estimation. ArXiv preprint. Available at arXiv:1309.5936.
[54] Wu, W., Olhede, S.C. and Wolfe, P.J. (2024). Supplement to "Tractably modelling dependence in networks beyond exchangeability." https://doi.org/10.3150/24-BEJ1740SUPP
[55] Wu, W. and Zhou, Z. (2017). Nonparametric inference for time-varying coefficient quantile regression. *J. Bus. Econom. Statist.* **35** 98–109. MR3591532 https://doi.org/10.1080/07350015.2015.1060884
[56] Wu, W.B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. USA* **102** 14150–14154. MR2172215 https://doi.org/10.1073/pnas.0506715102
[57] Yuan, Y. and Qu, A. (2021). Community detection with dependent connectivity. *Ann. Statist.* **49** 2378–2428. MR4319255 https://doi.org/10.1214/20-aos2042
[58] Zhang, Y., Levina, E. and Zhu, J. (2020). Detecting overlapping communities in networks using spectral methods. *SIAM J. Math. Data Sci.* **2** 265–283. MR4081803 https://doi.org/10.1137/19M1272238
[59] Zhao, Y., Levina, E. and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. MR3059083 https://doi.org/10.1214/12-AOS1036
[60] Zhou, Z. (2013). Heteroscedasticity and autocorrelation robust structural change detection. *J. Amer. Statist. Assoc.* **108** 726–740. MR3174655 https://doi.org/10.1080/01621459.2013.787184

[61] Zhou, Z. (2014). Inference of weighted V-statistics for nonstationary time series and its applications. *Ann. Statist.* **42** 87–114. MR3161462 https://doi.org/10.1214/13-AOS1184

[62] Zhou, Z. and Wu, W.B. (2010). Simultaneous inference of linear models with time varying coefficients. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 513–531. MR2758526 https://doi.org/10.1111/j.1467-9868.2010.00743.x

Received July 2022 and revised February 2024