Histogram-based approach for graphon estimation via joint exploitation of multiple networks

Roland B. Sogan and Tabea Rebafka

Sorbonne Université, Université Paris Cité, CNRS
Laboratoire de Probabilités, Statistique et Modélisation
4 place Jussieu, 75005 Paris, France
Email: roland-boniface.sogan@sorbonne-universite.fr,
tabea.rebafka@sorbonne-universite.fr

February 13, 2024

Abstract

Exchangeable graph models represent a commonly used non-parametric approach in the modeling of network data. They are characterized by a mathematical object called a graphon. This study focuses on estimating the graphon from multiple networks generated independently from the same model, with possibly distinct sets of nodes. We propose a new histogram estimator that leverages the joint sorting of empirical degrees of the graphs. Our algorithm is extremely fast and scalable to very huge datasets. A numerical study illustrates that the proposed estimator clearly outperforms naive estimators based on the average of individual graphon estimates. Our estimator is consistent when the number of nodes per network increases.

Keywords: Graphon, random graphs, multiple networks, histogram estimate.

1 Introduction

Machine learning on network-valued data is an active research field. For a long time research has focused on the analysis of a single network, but today in more and more applications entire sets of networks are observed. Thus, there is an increasing interest in the joint study of multiple networks. In this work, we consider the important problem of recovering the graphon from which a collection of graphs was generated. A graphon is a bi-variate function that represents the limiting structure of a sequence of graphs with increasing number of nodes (Lovász and Szegedy, 2006), but it can also be viewed as a nonparametric statistical model for exchangeable random graphs (Diaconis and Janson, 2007; Bickel and Chen, 2009). In general, the graphon is not uniquely defined, but only up to permutation of its parts. In some cases, a canonical representation of the graphon can be easily defined.

The focus of this work is on sets of networks without node correspondence, that is, every network comes with its own set of vertices. The lack of node correspondence represents a challenge for the analysis of the data, as in every graph node labels are arbitrary and can be

permuted in any way. Thus, comparing networks becomes tricky and cannot be done in a straightforward way. For instance, the Euclidean distance between two adjacency matrices is meaningless as it depends on the order of the nodes. Moreover, as node sets are different from one network to another, networks may also have different numbers of nodes, so that adjacency matrices are of different sizes and not comparable at all. Thus, the joint analysis of multiple networks without node correspondence is a challenge in general and in particular when it comes to the estimation of the graphon. To the best of our knowledge, the only graphon estimator for multiple networks that have different sets of nodes provided by the literature is based on the stochastic block model (SBM), which is a model with a specific piecewise constant graphon. That is, the observed networks are considered as independent and identically distributed (iid) realizations of a SBM. For fitting a SBM to a collection of networks, Chabert-Liddell et al. (2023) propose a variational EM-algorithm, while Rebafka (2024) proposes a hierarchical greedy algorithm. Both methods have long computing time and are not scalable to very huge datasets. The purpose of this work is to propose and discuss new methods for the fast estimation of the graphon in the multiple network setting.

In the literature of machine learning, several graphon estimation procedures have been proposed based on the observation of a single network or multiple networks with node correspondence, that is, the observed graphs share the same node set. Besides the above mentioned SBM, Airoldi et al. (2013) propose an estimation procedure for a graphon, called the stochastic blockmodel approximation (SBA) algorithm, computed on multiple networks with node correspondence. The basic idea of SBA is to approximate the graphon by a two-dimensional step function, which corresponds to a SBM. Now, SBA is a fast greedy algorithm, where the number of blocks is chosen in a data-driven way. So, when the number of nodes increases, a SBM with more blocks is selected yielding a step-function approximation of the graphon with ever finer intervals. As a consequence consistency of the SBA algorithm is obtained. Another approach is the universal singular value thresholding (USVT) algorithm for the observation of a single network (Chatterjee, 2012), which is based on a SVD of the adjacency matrix. The estimator is shown to be consistent when the number of nodes tends to infinity. Moreover, Lloyd et al. (2012) introduce a Bayesian nonparametric model by placing a Gaussian process prior on the graphon. However, there is no consistency guarantee of the estimator. None of these graphon estimators has canonical form. However, Chan and Airoldi (2014) propose a consistent and numerically efficient histogram estimator of the canonical form of the graphon. Their estimator is based on the so-called sorting and smoothing (SAS) algorithm, which first computes a histogram estimate of the graphon, which is then smoothed by some total variation minimization.

Generally, a simple way to derive an estimator for multiple networks consists in computing a graphon estimator on every network and then taking the average of all graphon estimates. However, this approach only makes sense when all individual graphon estimates have a canonical form. Besides considering the average of individual graphon estimates, none of the existing algorithms mentioned so far is appropriate to directly analyze multiple networks without node correspondence, since the joint analysis of several graphs is involved.

Contributions. We make two contributions in this paper. First, for any exchangeable graph model satisfying some identifiability condition, we introduce a novel and efficient method for estimating the graphon on multiple networks with different node sets. The proposed algorithm outperforms the approach where one applies a state-of-the-art algorithm to each of the networks and then takes their average. Second, the proposed estimator demonstrates superior performance, especially as the number of graphs tends to infinity. Notably, there is a significant improvement over existing state-of-the-art algorithms, which fix the number of graphs and where the network size tends to infinity. From this perspective, the proposed algorithm is

computationally more efficient compared to the greedy algorithm proposed by Rebafka (2024) on SBMs.

The rest of this work is organized as follows. In Section 2, we establish the framework of the study and discuss the graphon identifiability issue. In Section 3, we introduce and discuss the new estimator. Finally, in Section 4, we present the simulation study to assess its performance and compare to the state of the art.

2 Model

We observe a collection of networks $\mathcal{G} = (G^{(1)}, ..., G^{(M)})$, where the m-th graph $G^{(m)} = (V^{(m)}, E^{(m)})$ has node set $V^{(m)}$ and edge set $E^{(m)}$. We assume that the graphs are binary, undirected and do not have the same node sets $V^{(m)}$. The numbers of nodes $n_m = |V^{(m)}|$ may not be identical either. We denote the adjacency matrix of graph $G^{(m)}$ by $A^{(m)} \in \{0,1\}^{n_m \times n_m}$.

Let $\mathcal{G}$ be a set of independent and identically distributed exchangeable random graphs generated from some unknown graphon w. A **graphon** w is a symmetric measurable function $w : [0,1]^2 \to [0,1]$, where $w(u,v)$ represents the probability of an edge between nodes of the graph. For each m, $G^{(m)}$ is generated by the following sampling scheme. First, generate a uniformly distributed latent variable $U_i^{(m)}$ for each node $i \in \{1, ..., n_m\}$, that is,

$$U_1^{(m)}, ..., U_{n_m}^{(m)} \overset{iid}{\sim} \text{Uniform}[0,1].$$

Then, conditionally to these latent variables, the entries of adjacency matrix $A^{(m)}$ are generated from a Bernoulli distribution with parameter given by the graphon w. More precisely,

$$A_{ij}^{(m)} | U_i^{(m)}, U_j^{(m)} \overset{ind}{\sim} \text{Bernoulli}(w(U_i^{(m)}, U_j^{(m)})) \quad \text{for } i \leq j.$$

The goal is to estimate the graphon w from the data $\mathcal{G}$. For the estimation problem to be well-posed, we focus on identifiable graphons as defined below.

Condition 2.1 (**Strict monotonicity of degrees**). *A graphon w has a unique representation if there exists a measure-preserving transformation φ such that $w^{can}(u,v) = w(\varphi(u), \varphi(v))$ and*

$$g^{can}(u) = \int_0^1 w^{can}(u,v)dv,$$

*is strictly increasing. The graphon w^{can} is called the **canonical representation** of w.*

In the rest of this work, we mainly focus on graphons satisfying the strict monotonicity condition. For notational simplicity, we denote w as the canonical representation rather than w^{can}.

3 Joint Graph Sorting algorithm

The joint graph sorting (JGS) algorithm exploits the fact that the canonical graphon is identifiable and that the latent positions $U_i^{(m)}$ may be recovered by sorting the empirical degrees of the graphs. Then, using the estimated latent positions, a histogram estimate of the graphon is easily defined. This approach is used by several authors, such as Chan and Airoldi (2014), but by performing the sorting network by network. Here, we propose an estimation

procedure based on a joint analysis of the networks. Namely, we consider ordering the nodes not networkwisely, but establishing the latent position of each node with respect to the entire collection of networks. This improves the estimation of the latent variables, and consequently the estimation of the graphon.

Joint sorting stage: We compute the empirical degrees for all nodes of all the graphs. In order to put them on the same scale, we divide the empirical degrees by the number of nodes in the graph, and we refer to them as the **normalized empirical degrees**. Formally, we compute for $m = 1, ..., M$,

$$d_i^{(m)} = \frac{1}{n_m} \sum_{j=1}^{n_m} A_{i,j}^{(m)}, \text{ for } i = 1, ..., n_m.$$

Then, we consider the set of all nodes $V = \cup_{m=1}^{M} V^{(m)}$ and order them according to their normalized empirical degrees. More precisely, let $d^{(m)} = (d_1^{(m)}, ..., d_{n_m}^{(m)}) \in [0,1]^{n_m}$ be the sequence of normalized degrees of $G^{(m)}$, for $m = 1, ..., M$. Then we relabel the nodes in the following way. The nodes of the first graph $G^{(1)}$ are denoted by $\{1, ..., n_1\}$. Those of the second graph $G^{(2)}$ are denoted by $\{n_1 + 1, ..., n_1 + n_2\}$ and so on. Now, let $d = (d_1, ..., d_n) = (d_1^{(1)}, ..., d_{n_1}^{(1)}, ..., d_1^{(M)}, ..., d_{n_M}^{(M)})$ be the vector representing the normalized degrees of the nodes of all M graphs, where

$$n = \sum_{m=1}^{M} n_m = |V| = \sum_{m=1}^{M} |V^{(m)}|.$$

To recover the canonical graphon, we search a permutation $\hat{\sigma}$ of the nodes $\{1, ..., n\}$ such that $d_{\hat{\sigma}(1)} \leq \cdots \leq d_{\hat{\sigma}(n)}$. Ordering the nodes according to $\widehat{\sigma}$ amounts to estimate the latent positions by

$$\widehat{U}_i = \frac{\widehat{\sigma}^{-1}(i)}{n} - \frac{1}{2n}, \quad i = 1, ..., n. \tag{3.1}$$

Thus, compared to the previous graphon estimate, a finer grid for the values of the latent positions is considered.

Now, to compute a histogram estimate, we divide the interval $[0,1]$ in k regular intervals of length $h = \frac{1}{k}$ and compute the edge frequencies per block. Formally, let $J_l^{(m)} \subset \{1, ..., n_m\}$ be the set of node indices i_m of graph $G_{(m)}$ such that $\widehat{U}_i = \widehat{U}_{i_m}^{(m)} \in I_l = [(l-1)h, lh[$. Then, the edge frequency of block $I_s \times I_t$, $s, t = 1, ..., k$ is giving by

$$\widehat{H}_{s,t} = \frac{\sum_{m=1}^{M} \sum_{(i,j) \in J_s^{(m)} \times J_t^{(m)}} \widehat{A}_{ij}^{(m)}}{\max\left\{\sum_{m=1}^{M} |J_s^{(m)}| \cdot |J_t^{(m)}|, 1\right\}}. \tag{3.2}$$

Finally, the graphon estimate is giving by

$$\widehat{w}(u, v) = \widehat{H}_{s,t}, \ \forall (u, v) \in I_s \times I_t.$$

This estimation procedure is summarized in Algorithm 1. Compared to the methods by Rebafka (2024) and Chabert-Liddell et al. (2023), which fit a SBM to a collection of networks by using cumbersome iterative algorithms, our estimator is extremely fast as it consists of only two steps. It is also scalable to very huge datasets with a large number of networks and/or large network sizes.

Algorithm 1: Joint Graph Sorting algorithm

Input	: A collection of observed adjacency matrices $A^{(1)}, ..., A^{(M)}$ of respective sizes $n_1, ..., n_M$ and the desired number of histogram blocks k
Joint stage	: Compute the normalized empirical degrees $d_i^{(m)} = \frac{1}{n^{(m)}} \sum_{j=1}^{n^{(m)}} A_{i,j}^{(m)}$, $i = 1, ..., n_m, m = 1, ..., M$; Order all nodes according to their normalized degrees ; Compute the partition $(I_1, ..., I_k)$, and the sets $(J_1^{(m)}, ..., J_k^{(m)})$;
Global estimate:	For $s, t \in \{1, ..., k\}$, compute $\widehat{H}_{s,t}$ according to Equation (3.2)
Output	: Graphon estimate in form of matrix $\widehat{H} \in [0, 1]^{k \times k}$.

4 Simulations study

In this section, we illustrate the performance of the proposed estimator for the continuous graphon $w(u, v) = uv$ that satisfies Condition 2.1 on the monotonicity of the degree sequence.

We compare our estimator to the one which is obtained by computing the SAS estimator by Chan and Airoldi (2014) on each network and then taking the mean of the individual estimators. That is, the node sorting is done networkwisely, and not on the level of the entire set of networks as far as our estimator. Graphs were generated based on this graphon, and we use the mean integrated squared error (MISE) to compare our estimator to the SAS estimator of Chan and Airoldi (2014) in two settings. In the first one, the number $M = 20$ of graphs within each collection is fixed and the graph size n increases from 50 to 1000 (Figure 1.b). We remark that our JGS estimator is consistent and outperforms the SAS estimator. In the second setting, we vary the number M of graphs within each collection, while simultaneously keeping the number n of nodes in all graphs constant. More precisely, we generate collections where every graph has $n = 20$ nodes and the number M of networks increases from 17 to 500 (Figure 1.a). Several observations can be done. First, we see that as expected, the MISE of both estimators decreases when M increases. For every number of networks M, the proposed estimator largely outperforms the SAS estimator. Second, it is interesting to see that the MISE of both estimators does not vanish, indicating that none of the estimators is consistent. To understand this fact, it is instructive to analyze the estimators $\widehat{U}_i^{(m)}$ of the latent positions. Figure 2 represents the mean squared error (MSE) of the estimates $\widehat{U}_i^{(m)}$ for the two estimation procedures. In Fig 2.b we see that the latent positions are consistently estimated when adding nodes to each network. However, this is not the case in the second asymptotic setting (see Fig 2.a and its zoom). Indeed, adding now networks to the collection does not improve the normalize node degree $d_i^{(m)}$ as they are only computed on the adjacency matrix $A^{(m)}$ which remains fixed.

5 Conclusion

We have proposed an approach for nonparametric graphon estimation based on histograms in the multiple network setting without correspondence of the node sets. Estimating the graphon in this context is a challenging task, primarily due to the problem of graphon non-identifiability and the intricacy of network comparison. Our estimator is very fast and has significantly better accuracy than the SAS estimator. This improvement is due to a better exploitation of the data. Nevertheless, our estimator suffers from some limitations in the asymptotic setting where the numbers of networks increases, while the network sizes are bounded. In

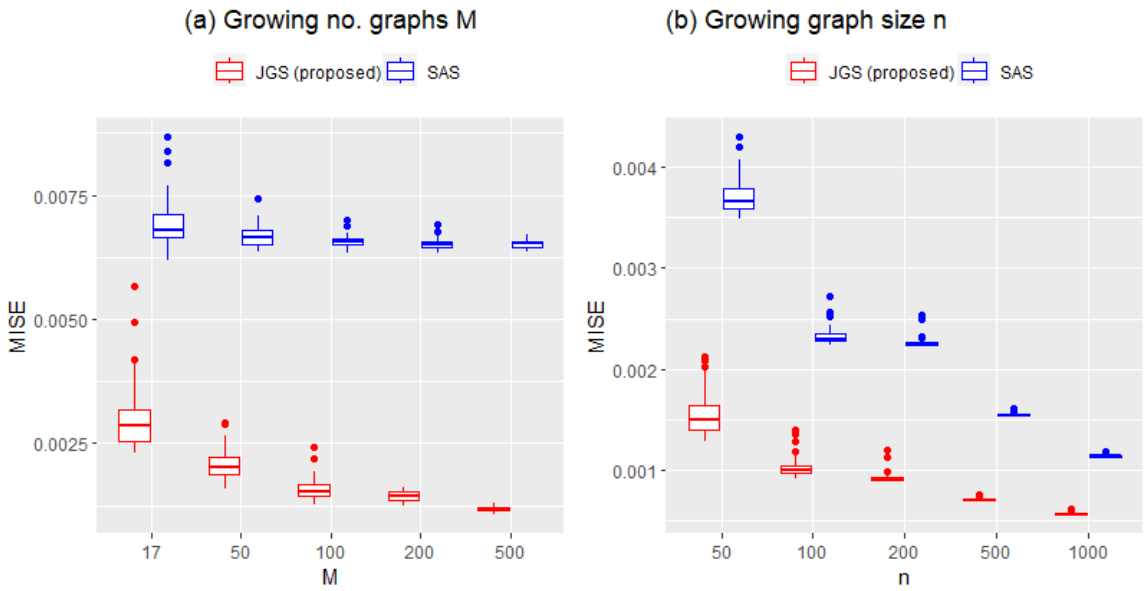

Figure 1: MISE of proposed (JGS) alogrithm vs MISE of SAS algorithm

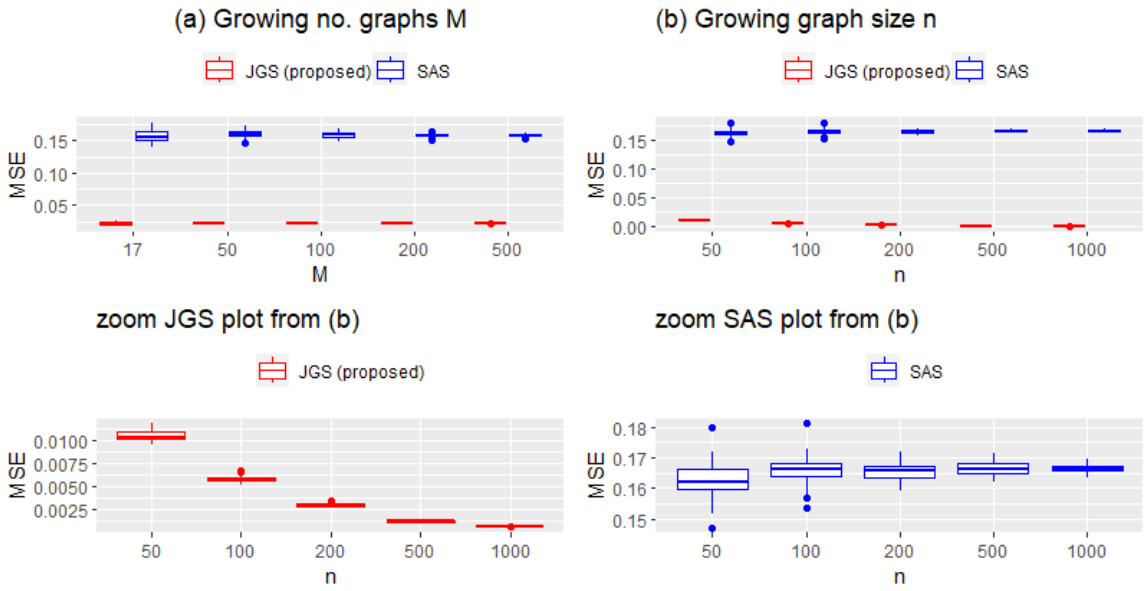

Figure 2: MSE of latent positions: the proposed (JGS) algorithm vs the SAS algorithm

this setting the data do not provide sufficiently much information to consistently estimate the latent node positions.

References

Airoldi, E. M., Costa, T. B., and Chan, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26, pages 692–700.

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(50):21068–21073.

Chabert-Liddell, S.-C., Barbillon, P., and Donnet, S. (2023). Learning common structures in a collection of networks: An application to food webs. ArXiv:2206.00560, Mar.

Chan, S. and Airoldi, E. (2014). A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning*, pages 208–216.

Chatterjee, S. (2012). Matrix estimation by universal singular value thresholding. ArXiv:1212.1247.

Diaconis, P. and Janson, S. (2007). Graph limits and exchangeable random graphs. *Rendiconti di Matematica e delle sue Applicazioni, Series VII*, pages 33–61.

Lloyd, J. R., Orbanz, P., Ghahramani, Z., and Roy, D. M. (2012). Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems (NIPS)*, volume 25, pages 1007–1015.

Lovász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96:933–957.

Rebafka, T. (2024). Model-based clustering of multiple networks with a hierarchical algorithm. *Statistics and Computing*, 34:1–16.