Nonparametric estimation and testing of exchangeable graph models

Justin J. Yang
Harvard University

Qiuyi Han
Harvard University

Edoardo M. Airoldi
Harvard University

Abstract

Exchangeable graph models (ExGM) are a nonparametric approach to modeling network data that subsumes a number of popular models. The key object that defines an ExGM is often referred to as a *graphon*, or *graph kernel*. Here, we make three contributions to advance the theory of estimation of graphons. We determine conditions under which a *unique* canonical representation for a graphon exists and it is identifiable. We propose a 3-step procedure to estimate the canonical graphon of any ExGM that satisfies these conditions. We then focus on a specific estimator, built using the proposed 3-step procedure, which combines probability matrix estimation by Universal Singular Value Thresholding (USVT) and empirical degree sorting of the observed adjacency matrix. We prove that this estimator is consistent. We illustrate how the proposed theory and methods can be used to develop hypothesis testing procedures for models of network data.

1 INTRODUCTION

Network data are ubiquitous and research approaches to network modeling have been gaining momentum in the past few years (Goldenberg et al., 2009; Kolaczyk, 2009). Applications span a diverse range of scientific areas, from biology and genetics, to social sciences, economics, and information sciences. These applications raise challenging questions about modeling, inference and computation.

Here, we focus on exchangeable graph models (ExGM; Aldous, 1981; Hoover, 1979; Kallenberg, 1989, 2005) and discuss circumstances under which the graphons that define them can be consistently estimated. A traditional way to formulate this estimation problem is to focus on the probability matrix, which generates the observed adjacency matrix in the sense of independent Bernoulli trials. Several recent papers focus on this direction (Bickel and Chen, 2009; Miller, Griffiths, and Jordan, 2009; Lloyd, Orbanz, Ghahramani, and Roy, 2012; Choi, Wolfe, and Airoldi, 2012; Azari and Airoldi, 2012; Chatterjee, 2012; Tang, Sussman, and Priebe, 2013; Wolfe and Olhede, 2013; Latouche and Robin, 2013; Orbanz and Roy, 2013; Airoldi, Costa, and Chan, 2013; Chan, Costa, and Airoldi, 2013; Chan and Airoldi, 2014), but one of the deficiencies for this formulation is that the resulting estimate always lacks the global structural information to the generating graphon.

In this paper, we adopt an alternative way to formulate the estimation problem for the generating graphon underlying an ExGM. Our goal in this work is to adopt, an operate within the constraints of, a fully functional form for the unknown graphon, and develop a nonparametric estimation strategy for the graphon that defines such an ExGM.

Contributions. We make three contributions in this paper. First, we clearly discuss identifiability issues, when pursuing a functional form estimation to the unknown graphon. In other words, before we have a functional form estimate, we need to uniquely define an estimand that is also in a functional form. We formalize an identifiability condition that requires an ExGM to have an absolutely continuous degree proportion distribution. A similar condition was originally implicitly leveraged by by Bickel and Chen (2009) in a different form, which we explicitly state and reformulate in our context. Under this condition, there is always an uniquely defined *canonical* representation for the generating graphons of an ExGM, which is the primary estimand of interest in this work.

Second, for any ExGM satisfying the identifiability condition, we propose a 3-step procedure to construct a flexible set of nonparametric estimates of the canonical graphon. This procedure requires (i) a latent variables estimation step, by empirical degree sorting, (ii) a probability matrix estimation step, and finally (iii) an optional smoothing step. This 3-step procedure allows the combination of any two strategies for latent variable and probability matrix estimation with a smoothing method. Leveraging this procedure, researchers can flexibly design nonparametric estimates of the canonical graphon, depending on their goals or willingness to state assumptions about the quantities involved.

Third, we consider a specific estimator, designed according

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

to the proposed 3-step procedure, by combining probability matrix estimation by Universal Singular Value Thresholding (USVT; Chatterjee, 2012) and empirical degree sorting using the observed adjacency matrix. This combination, which we refer to informally as the USVT-A procedure, is proved to be consistent, only requiring continuity on the true canonical graphon. The proposed USVT-A procedure is both computationally efficient and easy to implement. This makes simulation-based hypothesis testing of models of network data practical.

Outline. The rest of this paper is organized as follows. In Section 2, we discuss in details the identifiability issue to propose a functional form estimate for the generating graphon. In Section 3, we propose the 3-step procedure to construct estimates of a graphon. In Section 4, we focus on a specific choice of estimator and derive its theoretical asymptotic properties. In Section 5, we demonstrate the power of pursuing a functional form estimate in the context of classical hypothesis testing. We offer some remarks in Section 6.

2 IDENTIFIABILITY OF ExGM

Here we define some key notions. We then move on to the discussion of the identifiability of graphons and conclude with a special but flexible subclass of ExGM, which we will focus on in the remainder of the paper.

2.1 Basic setup

Let $U_1, ..., U_N$ be i.i.d. uniform random variables on the closed interval $[0,1]$, and let $W : [0,1]^2 \to [0,1]$ be an unknown symmetric measurable function. The observed data is an undirected simple graph described by an adjacent matrix A, which is a $N \times N$ symmetric random matrix with binary elements such that, for $\mathcal{U}_N \triangleq \sigma(U_1, ..., U_N)$,

$$\begin{aligned} A_{ii} &= 0 \text{ for each } i \text{ and} \\ A_{ij}|\mathcal{U}_N &\backsim \text{Ber}(W(U_i, U_j)) \text{ for } i < j, \end{aligned}$$

where A_{ij}'s are, conditionally on $\mathcal{U}_N$, independent to each other for $i < j$. The unknown symmetric parameter matrix

$$\begin{aligned} P_{ii} &\triangleq 0 \text{ for each } i \text{ and} \\ P_{ij} &\triangleq W(U_i, U_j) \text{ for } i < j \end{aligned}$$

is then called the probability matrix. This model specification on the observed undirected graph is then called the exchangeable graph model (ExGM), W is called a graphon generating this ExGM, and $U_1, ..., U_N$ are called the latent variables for the observed graph.

The goal is to draw inferences about the unknown graphon W from the observed adjacency matrix A. Researchers typically formulate the estimation problem as follows:

Estimation problem 1 (P1). Build a matrix estimator $\hat{P}$ of P, independently of any latent variable formulation.

Even though this is the most common way to formulate the estimation problem, this approach often leads to an estimator $\hat{P}$ that is unable to describe structural information encoded by the generating graphon W, which then blocks us from doing inferences on some interesting and practical problems, like model similarity checking or prediction inferences. Here, we pursue an alternative formulation of the estimation problem as follows:

Estimation problem 2 (P2). Build a nonparametric estimator $\hat{W}(u,v) \equiv W(\hat{u}, \hat{v})$ of $W(u,v)$, as a plug-in estimator that relies on estimating latent variables.

However, there is an unavoidable well-posedness issues before we further study estimation problem no. 2. Due to the highly symmetry structure resulted by the exchangeability of ExGM, several graphons might generate the same ExGM simultaneously, so estimation problem no. 2 won't be well-posed unless we can assign a unique and identifiable representation among those graphons generating the same underlying ExGM. We say two ExGMs $\mathbb{P}_1$ and $\mathbb{P}_2$ are the same if, for any binary and symmetric $N \times N$ matrix realization A and any $N \in \mathbb{N}$, $\mathbb{P}_1(A) = \mathbb{P}_2(A)$. We discuss this issue next.

2.2 Identifiability of graphons

The discussion of the identifiability for estimation problem no. 2 starts from a non-trivial statement, which seems to be true at first glance, for any two graphons generating the same ExGM. It is often stated that, for any measure preserving mapping $\varphi : [0,1] \to [0,1]$,

$$W'(u,v) \triangleq W(\varphi(u), \varphi(v)) \tag{1}$$

for almost everywhere[1] (a.e.) $(u,v) \in [0,1]^2$, generates the same ExGM as W. Conversely, for a given ExGM $\mathbb{P}$, is the relationship above the only uncertainty about W? In other words, suppose that both W and W' generate the same ExGM $\mathbb{P}$, does there exist a measure preserving mapping φ such that equation (1) holds?

The answer is negative, while the opposite wrong answer has been often misused in the statistical literature about ExGM. An intuitive counterexample proposed by Diaconis and Janson (2008) is as follows,

$$\begin{aligned} W(u,v) &= uv \text{ and} \\ W'(u,v) &\triangleq (2u \bmod 1)(2v \bmod 1). \end{aligned}$$

Then these two graphons will generate the same ExGM but there exists no such a measure preserving mapping φ satisfying equation (1). A more accurate condition for identifiability is given in Theorem 7.1 by Diaconis and Janson

[1] For $[0,1]^d$ space, we always refer the term almost everywhere with respect to the complete Lebesgue measure on it.

(2008), which states that W and W' will generate the same ExGM if and only if there exist two—rather than one—measure preserving mappings φ and φ' such that

$$W(\varphi(u), \varphi(v)) = W'(\varphi'(u), \varphi'(v)) \text{ a.e. } (u,v) \in [0,1]^2.$$

Note that an alternative and equivalent characterization that ensures W and W' generate the same ExGM is

$$\delta_{\square}(W, W') = 0,$$

where $\delta_{\square}$ is the so-called cut-metric (Borgs et al., 2008).

However, the result by Diaconis and Janson (2008) doesn't mean that equation (1) should be fully abandoned. For such a relationship to hold among graphons generating the same ExGM, the following condition must be satisfied:

Definition 1 (Twin-free condition) *There exists no such a pair* (u_1, u_2) *in* $[0,1]$ *such that* $W(u_1, v) = W(u_2, v)$ *for a.e.* $v \in [0,1]$.

For any two twin-free graphons W_1 and W_2 generating the same ExGM, Borgs, Chayes, and Lovász (2010) proved that there is actually a measure preserving bijection $\varphi_{12} : [0,1] \to [0,1]$ such that

$$W_1(u,v) = W_2(\varphi_{12}(u), \varphi_{12}(v)) \text{ for a.e. } (u,v) \in [0,1]^2.$$

Thus, for those papers misusing equation (1) as the only condition to define graphons generating the same ExGM, a simple fix would be to rephrase the results by limiting their interests to a subclass of ExGMs generated by twin-free graphons, which we call *twin-freely* generated ExGMs. Unfortunately, as of this writing, there is no known result that states an appropriate way to choose a unique representation for graphons that generate a twin-free ExGM.

In oder to resolve this issue, in the rest of this paper we will consider a relatively more restrictive subclass of ExGM, following and expanding on the seminal paper by Bickel and Chen (2009). They attempt to solve the identifiability issue by claiming that, for any ExGM $\mathbb{P}$ generated by a graphon W, one can find a measure preserving mapping φ such that, for $W^{\mathbb{P}}_{\text{can}} \triangleq W(\varphi(u), \varphi(v))$,

$$g^{\mathbb{P}}_{\text{can}}(u) \triangleq \int_0^1 W^{\mathbb{P}}_{\text{can}}(u,v)\, dv$$

is monotone non-decreasing for $u \in [0,1]$. They also argued that the so-called canonical form $W^{\mathbb{P}}_{\text{can}}$ of the graphon W is uniquely determined for a.e. $(u,v) \in [0,1]^2$.

We expand their condition by assuming the following:

Definition 2 (Degree-identifiability condition) *Let* U *be a uniform random variable on* $[0,1]$. *Then the degree proportion*

$$g(U) \triangleq \int_0^1 W(U,v)\, dv$$

is an absolutely continuous random variable[2] *on* $[0,1]$.

In later work, Bickel et al. (2011) also rely on a similar assumption to ensure identifiability of the graphon.

An easy example to check why this extension is necessary is given by the following two graphons

$$\begin{aligned} W(u,v) &\triangleq 1_{[0,1/2]^2}(u,v) + 1_{[1/2,1]^2}(u,v), \\ W'(u,v) &\triangleq 1_{[0,1/2]\times[1/2,1]}(u,v) + 1_{[1/2,1]\times[0,1/2]}(u,v), \end{aligned}$$

which give monotone non-decreasing $g(u) \equiv g'(u) \equiv 1/2$, generate a same ExGM, yet are different for a.e. $(u,v) \in [0,1]^2$. There is no canonical choice between W and W' in this example.

Therefore, estimation problem no. 2 stated above will be well-posed as long as we focus on a subclass of ExGMs generated by degree-identifiable graphons, and if we treat the uniquely defined canonical graphon $W^{\mathbb{P}}_{\text{can}}$ associated with a degree-identifiable ExGM $\mathbb{P}$ as the major estimand of interest. The next section will discuss a strategy to develop estimation procedures in this context.

Remark 1 *Starting from the next Section, we will simply write* W *and* g *to refer the canonical graphon of a degree-identifiable ExGM and its marginal integral.*

Remark 2 *There are actually three equivalent characterizations for a degree-identifiable ExGM* $\mathbb{P}$*:*

1. $g(U)$ *is an absolutely continuous random variable;*
2. $g^{\mathbb{P}}_{\text{can}}$ *is strictly increasing on* $[0,1]$.
3. *The cumulative distribution function (CDF) of* $g(U)$ *is absolutely continuous and hence is continuous.*

3 THREE-STEP ESTIMATION OF DEGREE-IDENTIFIABLE ExGMs

In this Section, we will explain how, in a three steps procedure, to construct a flexible class of functional form or nonparametric estimates for the canonical graphon generating a degree-identifiable ExGM. Then we will conclude with a special choice of nonparametric estimate.

The main idea behind the estimation procedure is to exploit the degree-identifiable feature of the canonical graphon and make use of empirical degree sorting to infer unknown latent variables. We now describe how to proceed this three steps procedure in the following paragraphs.

Step 1: Probability matrix estimation. Perform any P1 estimation $\hat{P}$ for the probability matrix P.

[2]We should note that the random variable $g(U)$ here is uniquely determined by the ExGM $\mathbb{P}$ in the distribution sense.

Step 2: Latent variables estimation. Construct an empirical CDF of degree proportions using another P1 estimation $\hat{P}'$, which may or may not be the same as $\hat{P}$, and then let $\hat{U}_i$'s be the estimators of the unknown latent variables U_i's defined as the values of the empirical CDF evaluating at the degree proportions of i-th node in $\hat{P}'$.

The rationale of doing this Step 2 is explained here. According to some simulation evidences, the empirical CDF $\hat{F}(x)$ of degree proportions seems to describe the CDF of $g(U)$, which we denote as $g^{-1}(x)$, quite accurately when the number of nodes N is large enough. On the other hand, the law of large numbers can somehow guarantee that the degree proportions in P' at i-th node, $\frac{1}{N}\sum_{j=1}^{N} P'_{ij}$, will be a good approximation to $g(U_i)$ (assuming that given U_i, P'_{ij}'s are roughly i.i.d. from the distribution $W(U_i, U)$). As for a degree-identifiable ExGM, which requires the canonical graphon marginal integral g to be strictly increasing, we must have $u = g^{-1}(g(u))$ for every $u \in [0,1]$, so we can trust the estimation of the latent variables $U_i = g^{-1}(g(U_i))$ by $\hat{U}_i \triangleq \hat{F}\left(\frac{1}{N}\sum_{j=1}^{N} P'_{ij}\right)$.

Remark 3 *In the descriptions above, we temporarily assume that there is no overlapping for degree proportions in $\hat{P}'$, that is, we assume that the elements of*

$$\left\{\frac{1}{N}\sum_{j=1}^{N} P'_{ij}\right\}_{i=1}^{N}$$

are distinct. We will resolve this overlapping issue later after we commit to a specific choice for $\hat{P}'$.

Once we have conducted Step 1 and 2, we can start to construct a functional form estimate $\hat{W}(u,v)$. For now, we already have a set of three dimensional points

$$\left(\hat{U}_i, \hat{U}_j, \hat{W}\left(\hat{U}_i, \hat{U}_j\right)\right) \triangleq \left(\hat{U}_i, \hat{U}_j, \hat{P}_{ij}\right),$$

which we should treat as a noisy realization[3] of the unknown canonical graphon plane at $(U_i, U_j, W(U_i, U_j))$. To build a functional form estimation $\hat{W}(u,v)$ from those three dimensional points, we can either use a linear interpolation or a stepwise approximation as the pre-smoothing estimate. We majorly focus on the later one in this study, so the pre-smoothing estimate now takes the form of a step function

$$\hat{W}(u,v) \triangleq \sum_{1\le i,j\le N} \hat{P}_{ij} 1_{\left(\hat{U}_i - 1/N, \hat{U}_i\right] \times \left(\hat{U}_j - 1/N, \hat{U}_j\right]}(u,v).$$

Step 3: Smoothing. (Optional.) Apply any smoothing algorithm on the estimate to get a smoothed estimate, which may or may not be in a form of step function.

[3] With noise coming from not only the z-direction, but also the x- and y-directions as well.

Here are several notes related to this Step 3. First, it's a an optional step, and the choice of whether to include this step or not and how to conduct it depends on researchers' ultimate goal for inferences on network data and acceptability to those unavoidably additional assumptions on the canonical graphon. A detailed investigation of adding this third step in the estimation of canonical graphon is discussed in a follow-up paper (Chan and Airoldi, 2014).

Because both the Step 1 and 2 above can take any kind of estimator for problem no. 1 to proceed, we need to know how to explicitly specify $\hat{P}$ and $\hat{P}'$. Based on a comparative simulation study, which we omit for the sake of space, we select Universal Singular Value Thresholding (USVT) Chatterjee (2012) as the solution estimation problem no. 1 (in Step 1), and the adjacency matrix A itself as the basis for degree sorting (in Step 2). In the remainder of the paper we focus on this combination for estimation in order to seek the least assumptions on W; we refer informally to this method as the USVT-A estimation procedure.

3.1 Comparative Simulation Study

In this Subsection, we demonstrate two simulations showing the performance of different combinations of graphon estimations constructed from the 3-step procedure.

In each simulation, we calculate the root of the mean square error (RMSE) between the constructed estimator (using N ranging from 300 to 3000) and the true graphon, where only two cases are considered here: the quadratic graphon $W(u,v) = \left(u^2 + v^2\right)/4$ and the logistic graphon $W(u,v) = \text{logistic}(-5 + 5(u+v))$, where $\text{logistic}(x) \triangleq (1+\exp(-x))^{-1}$.

The simulation results are shown in Table 1 and 2.

In these Tables, the naming convention for each combination is to report the methods that were used for each of the steps separated by a dash; for example, "Step 1 method"-"Step 2 method"-"Step 3 method". The last smoothing step is optional–for example, USVT-A-TVM method stands for using USVT in Step 1 for probability matrix estimation, using the plain adjacency matrix A in Step 2 for the empirical degree sorting, and finally using the total variation smoothing (TVM) in Step 3. We also include the combination A-A as a baseline estimation procedure.

From the two Tables, we see that, for Step 1, using USVT is clearly better than using the vanilla A; for Step 2, sorting according to USVT estimate gives approximately the same result as (sometimes worse than) sorting according to the plain A; for Step 3, TVM smoothing can be helpful and reduce some mean square errors. It's interesting that A-A-TVM method gives a fairly good performance as both USVT-USVT-TVM and USVT-A-TVM methods. This observation motivated recent follow-up work (Chan and Airoldi, 2014).

While, adding a third smoothing step is helpful in these two

N	300	900	1500
A-A	0.344657	0.359469	0.357767
USVT-USVT	0.035505	0.024235	0.017006
USVT-A	0.037397	0.024614	0.017132
A-A-TVM	0.02453	0.013044	0.009418
USVT-USVT-TVM	0.040357	0.01202	0.008492
USVT-A-TVM	0.040601	0.011967	0.008526
N	1800	2400	3000
A-A	0.351921	0.360604	0.358457
USVT-USVT	0.01695	0.013922	0.012097
USVT-A	0.017059	0.013995	0.012168
A-A-TV	0.010491	0.00895	0.007025
USVT-USVT-TV	0.009912	0.00813	0.006128
USVT-A-TV	0.009926	0.008101	0.006066

Table 1: RMSE Simulation for Quadratic Graphon

N	300	900	1500
A-A	0.38003	0.3812	0.383409
USVT-USVT	0.102721	0.065759	0.034483
USVT-A	0.106061	0.069602	0.034192
A-A-TVM	0.085428	0.034208	0.02423
USVT-USVT-TVM	0.084122	0.051107	0.023318
USVT-A-TVM	0.075824	0.043535	0.02324
N	1800	2400	3000
A-A	0.380116	0.379474	0.379676
USVT-USVT	0.03164	0.024988	0.019176
USVT-A	0.031406	0.024771	0.019082
A-A-TVM	0.023198	0.019551	0.014232
USVT-USVT-TVM	0.023003	0.019187	0.014117
USVT-A-TVM	0.022963	0.019178	0.014113

Table 2: RMSE Simulation of Logistic Graphon

specific examples, we note that the *histogram estimator* recently proposed by Chan and Airoldi (2014) requires an additional smoothness assumption on the underlying canonical graphon W. The proposed that USVT-A estimation has only slightly larger RMSE than the histogram estimator, and its decreasing rate on the RMSE as the number of nodes increases is the same as that of the histogram estimator. Thus the proposed USVT-A method has a fairly good performance when compared with that of the histogram estimator, but its theoretical properties are achieved with less constraints on the graphon.

In the remainder of the paper, we focus on the USVT-A estimation in order to seek the least assumptions on W. Its theoretical property is discussed in the next Section.

4 CONSISTENCY

In this Section, we study the theoretical consistency of the USVT-A estimation procedure, which is defined through a combination of probability matrix estimation using USVT and the latent variables estimation using the empirical CDF of observed degrees proportions in A.

The main theoretical result of this paper is as follows:

Theorem 1 (USVT-A Consistency) *Assume that W is the canonical graphon of a degree-identifiable ExGM. If W is continuous on $[0,1]^2$, then the $\hat{W}$ constructed by the USVT-A method is consistent for estimating W in the sense that*

$$\mathbb{E}\left(\int_0^1\int_0^1\left(\hat{W}(u,v)-W(u,v)\right)^2 dudv\right)\to 0.$$

Here are two cornerstones that make our main result hold, both of which correspond to the consistency of Step 1 and Step 2 in our proposed three steps procedure in Section 3.

Theorem 2 (USVT Consistency) *Let*

$$\hat{M}\triangleq\sum_{i\ni\left\{s_i\geq 1.01\sqrt{N}\right\}} s_iu_iu_i^T \text{ and } \hat{P}_{ij}\triangleq\left(\left(\hat{M}_{ij}\right)\wedge 1\right)\vee 0,$$

be the USVT estimation of probability matrix P, where $\sum_{i=1}^N s_iu_iu_i^T$ is the singular value decomposition of adjacency matrix A. Then

$$\mathbb{E}\left(\frac{1}{N^2}\sum_{i,j=1}^N\left|\hat{P}_{ij}-P_{ij}\right|^2\right)\to 0. \tag{2}$$

See Chatterjee (2012) for a proof of the above theorem.

Here is some notation we are going to use throughout this Section. Let the observed degree proportions in A to be

$$D_i\triangleq\frac{1}{N}\sum_{j=1}^N A_{ij},$$

with their empirical CDF defined as

$$\hat{F}(x)\triangleq\frac{1}{N}\sum_{i=1}^N 1_{\{D_i\leq x\}}.$$

Theorem 3 (Degree Sorting Consistency) *Let the empirical degree sorting estimate of the latent variables to be*

$$\hat{U}_i\triangleq\hat{F}(D_i), \tag{3}$$

but to avoid the overlapping issues we will instead use

$$\tilde{U}_i\triangleq\hat{U}_i-\frac{\kappa_i-1}{N} \tag{4}$$

in the proposed USVT-A estimation, where κ_i, given all of $\hat{U}_i$, is jointly distributed as follows: let $C_1,...,C_M$ be those unique values of D_i's, and, if $D_{i_1}=\cdots=D_{i_{k_m}}=C_m$, then $\kappa_{i_1},...,\kappa_{i_{k_m}}$ are a uniform resampling of the set $\{1,2,...,k_m\}$. Then we have, for each $i=1,...,N$, $\left|U_i-\hat{U}_i\right|\to 0$ and $\left|\tilde{U}_i-\hat{U}_i\right|\to 0$ in probability, and hence $\left|U_i-\tilde{U}_i\right|\to 0$ in probability.

Theorem 3 describes the consistency of latent variables estimation via empirical degree sorting using the observed adjacency matrix A. For the sake clarity, a detailed proof of Theorem 3 is omitted, and can be found in a follow up report (Yang et al., 2014). A proof for Theorem 1 using the two key consistency results described above can also be found in a follow up report (Yang et al., 2014).

5 HYPOTHESIS TESTING

In this Section, we illustrate how the proposed USVT-A estimator can help to develop a classical hypothesis testing procedure for the analysis of network data. There is ample room for improvement of the procedure we describe here.

Hypothesis testing is a powerful procedure with limited literature in network data analysis. Olding and Wolfe (2009) presents likelihood ratio tests on three examples with a novel flavor: (i) Erdős-Rényi graph (Erdös and Renyi, 1959; Gilbert, 1959), (ii) stochastic blockmodel (Nowicki and Snijders, 2001; Airoldi et al., 2008), and (iii) the fixed-degree graph model (Blitzstein and Diaconis, 2006). However, such a method lacks the flexibility to cope with more sophisticated models, such as exchangeable graph models.

There are mainly two reasons why it is difficult to extend classical hypothesis testing theory to network data. The first is that modeling network data often involves latent variables. In case of ExGM, the U_i's are especially hard to handle. The second reason is the high computational cost of fitting existing methods, so it is often untenable to get the sampling distribution of the test statistic under the null hypothesis, using simulations. Instead, the proposed USVT-A procedure captures the structure in exchangeable graph models by design and is computationally efficient, so that Monte Carlo can be employed for obtaining the sampling distribution. A simple illustrative example follows.

Suppose that we observe network data represented by an adjacency matrix A, which is generated by a degree-identifiable ExGM with canonical graphon W. We want to test the two hypothesis: for $W_Q(u,v) \triangleq \frac{1}{4}(u^2+v^2)$,

$$H_0 : W(u,v) = W_Q(u,v) \text{ versus } H_a : W(u,v) \neq W_Q(u,v).$$

By Theorem 1, we will have the USVT-A estimate $\hat{W}$ is getting closer and closer in the sense of mean square errors to the true canonical graphon W when N is sufficiently large. Thus we can choose the test statistic to be

$$\begin{aligned} T &\triangleq \sqrt{\int_0^1 \int_0^1 \left| W_Q(u,v) - \hat{W}(u,v) \right|^2 du dv} \quad (5) \\ &\triangleq \left\| W_Q - \hat{W} \right\|, \end{aligned}$$

the L^2 distance between W_Q and $\hat{W}$. Although we can't analytically know the sampling distribution of T under the null hypothesis H_0, we can easily approximate it using a large amount of simulations because of the fast implementation of our USVT-A method. Using 5000 Monte Carlo samples for $N = 3000$, we get the histogram of T as in Figure 1. We can see that the sampling distribution of T under H_0 is right skewed. The mean and standard deviation of the Monte Carlo samples are 0.0115 and 5.656×10^{-4}, and the 95% quantile is 0.0126, so the rejection region is $T \geq 0.0126$.

Given an observed adjacency matrix A, we can calculate its corresponding $\hat{W}$ and T, and then reject H_0 if $T \geq 0.0126$.

6 DISCUSSION

Here we review some common ways to estimate the canonical graphon W underlying a degree-identifiable ExGM, and compare them to the proposed USVT-A estimator.

The first study of the properties of of an estimator for a graphon W has been carried out for a specific family of models; Bickel and Chen (2009) considered blockmodels with a fixed number K of blocks (Nowicki and Snijders, 2001), explicitly write down the underlying graphon, and show how the maximum likelihood estimator for the blockmodel matrix (which parametrizes the graphon) is consistent, while approaches based on modularity are not. Their estimator has some disadvantages: (i) the computational cost is high, although in line with other estimators for blockmodels, and (ii) the need to pre-specify a fixed number of blocks, K, introduces a difficult model selection issue, in practice. Even though these authors formulate the estimation problem as P2, defined in Section 2.1, they did restrict the estimation task to the parametric family of blockmodels, thus coming short of defining a general estimator $\hat{W}$ for the graphon W defining an ExGM.

To address some of the shortcomings in listed above, Rohe et al. (2011) and Choi et al. (2012) consider a generalized blockmodel with a growing number of classes, $K = O\left(N^{1/2}\right)$. In this setting, inference does not need to have K pre-specified, while at the same time empirically leads to smaller model bias when compared with Bickel and Chen (2009). However, these authors still consider a parametric family of blockmodels, although less restrictive, and the estimation task is computationally expensive.

Bickel et al. (2011) addresses the graphon estimation problem as a P2 formulation. These authors proposed a method of moment estimator that takes advantage of the counts of special subgraphs, referred to as *wheels*, in the observed network. They theoretically characterize the unknown graphon in terms of an abstract linear functional, based on the moments. While this approach is elegant, and leads to consistency in the absence of many assumptions on the graphon underlying an EcGM, the implied estimator is unfeasible, in practice. This happens because of the number of wheels is huge and counting the frequency of even a small subset of them is impractical. In addition, even

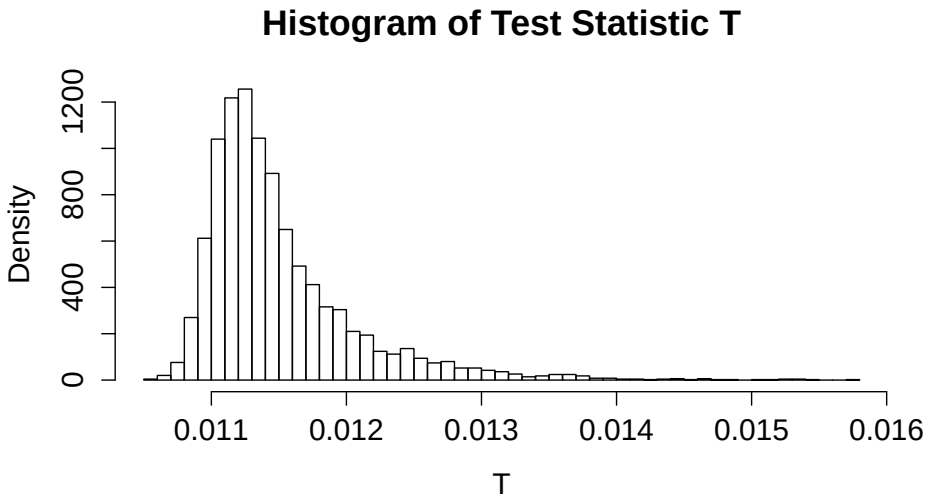

Figure 1: 5000 Monte Caro draws of T under H_0

given these counts, solving the canonical graphon from the characteristic linear functional described above is also challenging, because the need to compute the eigenvalues and eigenvectors of the characteristic functional.

More recent work by Wolfe and Olhede (2013) also develops a nonparametric estimator for a graphon underlying an ExGM. These authors measure the error between the true graphon and the estimated graphon via the cut-metric (Borgs et al., 2008). The estimator is thus defined implicitly by an equation that is solvable only in theory. Their results do not allow explicitly numerical simulations to check the performance of the estimation. In addition, the asymptotic theory requires sophisticated assumptions, which may be untenable in practice.

In contrast, the proposed USVT-A estimator is computationally efficient, easy to implement, and come with the same consistency guarantees of existing methods, with little assumptions on the underlying graphon.

6.1 Concluding remarks

In this paper, we have reformulated the existing literature on estimation problems for exchangeable graph models (ExGM), and dichotomized the existing approaches into two formulations; P1, addressing only on the probability matrix estimation, and P2, pursuing the fully functional form estimate for the graphon underlying an ExGM.

We discussed the important issue of identifiability, which must be addressed before any attempts on addressing the P2 formulation of the estimation problem can take place. We characterized a subclass of exchangeable graph models, referred to as degree-identifiable ExGMs, which entails a uniquely-defined marginal degree function for the canonical graphon, and leads to a well-posed estimation problem. Within this subclass of models, we proposed a general 3-step procedure for constructing a flexible class of nonparametric estimates of the canonical graphon, which allows a large number of combinations of (i) probability matrix estimation methods, (ii) latent variable estimation methods, and (iii) smoothing methods.

We then focused on a pre-smoothing estimate, which we refer to as the USVT-A method. We theoretically proved its mean square error consistency, under the only assumption of continuity of the canonical graphon degree function, which is easy to implement. Simulation results demonstrate the computational efficiency of the proposed USVT-A estimator, as well as its error properties, in practice. Our results also suggest that, if the canonical graphon W is believed to be smooth, then a smoothing algorithm like total variation minimization method (Chan, Khoshabeh, Gibson, Gill, and Nguyen, 2011) could be applied to get a further reduction of estimation errors (e.g., see Chan and Airoldi, 2014). However, simulation results also show that the reduction in RMSE obtained using total variation minimization seems to be relatively small. Other combinations of matrix estimators, latent variable estimators and smoothing methods should be considered as a promising avenue for future research.

References

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed-membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

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

D. J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11:581–598, December 1981.

H. Azari and E. M. Airoldi. Graphlet decomposition of a weighted network. *Journal of Machine Learning Research, W&CP*, 22:54–63, 2012.

P. J. Bickel and A. Chen. A nonparametric view of network models and newman-girvan and other modularities. In *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, 21068–21073, 2009.

P. J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *Annals of Statistics*, 39(5):2280–2301, 2011.

J. K. Blitzstein and P. Diaconis. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. Technical report, Stanford University, 2006.

C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi. Convergent sequences of dense graph I: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219:1801–1851, 2008.

C. Borgs, J. T. Chayes, and L. Lovász. Moments of two-variable functions and the uniqueness of graph limits. *Geometric and Functional Analysis*, 19:1597–1619, 2010.

S. H. Chan and E. M. Airoldi. A consistent histogram estimator for exchangeable graph models. *Journal of Machine Learning Research, W&CP*, 32:208–216, 2014.

S. H. Chan, R. Khoshabeh, K. B. Gibson, P. E. Gill, and T. Q. Nguyen. An augmented lagrangian method for total variation video restoration. *IEEE Transactions on Image Processing*, 20(11):3097–3111, 2011.

S. H. Chan, T. B. Costa, and E. M. Airoldi. Estimation of exchangeable random graph models by stochastic blockmodel approximation. In *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 293–296, 2013.

S. Chatterjee. Matrix estimation by universal singular value thresholding. ArXiv:1212.1247, 2012. Unpublished manuscript.

D. S. Choi, P. J. Wolfe, and E. M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99:273–284, 2012.

P. Diaconis and S. Janson. Graph limits and exchangeable random graphs. *Rendiconti di Matematica edelle sue Applicazioni, Series VII*, 28:33–61, 2008.

P. Erdös and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

E. N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2:129–233, 2009.

D. N. Hoover. Relations on probability spaces and arrays of random variables. Preprint, *Institute for Advanced Study*, Princeton, NJ, 1979.

O. Kallenberg. On the representation theorem for exchangeable arrays. *Journal of Multivariate Analysis*, 30 (1):137–154, 1989.

O. Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer, 2005.

E. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.

P. Latouche and S. Robin. Bayesian model averaging of stochastic block models to estimate the graphon function and motif frequencies in a w-graph model. ArXiv:1310.6150, 2013. Unpublished manuscript.

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

K. T. Miller, T. S. Griffiths, and M. I. Jordan. Nonparametric latent fature models for link prediction. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

K. Nowicki and T. Snijders. Estimation and prediction of stochastic block structures. *Journal of American Statistical Association*, 96:1077–1087, 2001.

B. P. Olding and P. J. Wolfe. Inference for graphs and networks: Extending classical tools to modern data. ArXiv:0906.4980, 2009. Unpulished manuscript.

P. Orbanz and D. M. Roy. Bayesian models of graphs, arrays and other exchangeable random structures, 2013. Unpublished manuscript.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

M. Tang, D. L. Sussman, and C. E. Priebe. Universally consistent vertex classification for latent positions graphs. *Annals of Statistics*, 2013. In press.

P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. ArXiv:1309.5936, 2013. Unpublished manuscript.

J. J. Yang, Q. Han, and E. M. Airoldi. Nonparametric estimation and testing of exchangeable graph models. Technical Report, 2014.