Edge-Induced Sampling from Graphons

Nicolas Kim & Alessandro Rinaldo
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
nicolask@stat.cmu.edu, arinaldo@cmu.edu

September 1, 2017

Abstract

The graphon model for random graphs is a general class of node-exchangable random graph models, which includes Stochastic Block Models and Latent Space Models as special cases. Various estimation procedures for graphons exist in the literature. However, accurate estimation requires either an unbiased sample or some way to account for potential bias in the sampling scheme. This presents a challenge for estimating modern social networks, where the edge sparsity can lead to unreliable estimates of the graphon based on the subgraph induced by a uniform sample of nodes. To alleviate this issue, we consider edge-induced subgraphs. To account for the sampling bias, we establish theory that describes the sampling distribution of these edge-induced subgraphs of graphons.

1 Introduction

Graphons were originally introduced as limiting objects for sequences of dense graphs by [18]. Graphons can be considered to be a vast generalization of the SBM, and in fact, any exchangeable random graph model is a graphon. In this section, we establish some basic theory for what graphons look like locally. More precisely, consider the graph model induced by taking only the nodes which are connected to a particular node in the network. In some sense, we are discussing a vertical (equivalently horizontal) slice of the whole-network graphon. The act of looking only at nodes which satisfy the additional constraint of being connected to some ego node should induce a *local graphon*, which is in general different from the original *ambient graphon*.

There is a rich literature for estimating the parameters of SBMs [12, 20] and graphons [1, 8, 23, 3]. These methods assume that the observed network was essentially obtained by observing the nodes uniformly at random. Comparatively little work has been done on estimating model parameters from an edge-wise sampling scheme compared to the node-wise case. A Bayesian procedure for

estimating the whole-network parameters for an SBM from a snowball sample is outlined in [21]. Work by [11, 15, 16, 22] establishes further results, with several simulated results, towards understanding the effect that snowball/respondent-driven sampling has on network models. To our knowledge, there are no such results in the literature for graphon estimation from edge-wise samples.

Section 2 establishes a theoretical framework for understanding the local structure of graphon models (of which the SBM is a special case). Section 3 introduces properties of snowball samples, and Section 4 discusses the estimation problem for snowball sampling.

2 Local Graphons

The graphon model defines the probability of an edge between nodes i and j as

$$\mathbb{P}\left(A_{ij} = 1 \mid U_i, U_j\right) = w(u_i, u_j),$$

where U_i and U_j are independent draws from $\text{Unif}(0, 1)$. These edge probabilities are then conditionally independent over the pairs of nodes given knowledge of the U's. This implies that

$$\mathbb{P}\left(A_{ij} = 1 \mid U_i\right) = \int_0^1 w(u_i, y)\,\mathrm{d}y.$$

This conditional tie probability is also known as the *degree function* [17], which is denoted by

$$d_w(x) \equiv \int_0^1 w(x, y)\,\mathrm{d}y.$$

Remark 1. *In the case of an SBM, this evaluates to*

$$\begin{aligned}
\mathbb{P}\left(A_{ij} = 1 \mid U_i\right) &= \int_0^{\alpha_1} \pi_{z_i,1}\,\mathrm{d}y + \int_{\alpha_1}^{\alpha_2+\alpha_1} \pi_{z_i,2}\,\mathrm{d}y + \cdots + \int_{1-\alpha_Q}^{1} \pi_{z_i,Q}\,\mathrm{d}y \\
&= \sum_{q=1}^{Q} \pi_{z_i,q}\alpha_q \\
&= \bar{\pi}_{z_i},
\end{aligned}$$

where $\bar{\pi}_{z_i}$ is equivalent to the symbol defined in Remark 1 of [13].

Corollary 1. *An implication of the degree function is that, since the edge probabilities are iid given U_i,*

$$\sum_{j \neq i} A_{ij} \mid U_i \sim \text{Bin}\left(n - 1, \int_0^1 w(u_i, y)\,\mathrm{d}y\right).$$

Hence, in a sample of n nodes from w, the expected degree of the node i with associated parameter x is $(n-1)\int_0^1 w(x, y)\,\mathrm{d}y$. In other words, nodes are entering the local (ego) network around the node i at the rate $\int_0^1 w(x, y)\,\mathrm{d}y$.

2.1 Filtering a Graphon through the Slice $w(U_i, -)$

There is a geometric intuition that the local graphon is induced by filtering the whole-network graphon through the slice $w(U_i, -)$. Looking at such a slice, one will notice regions of high values, which signify higher probabilities of sampling into the local network. The values of the graphon along the slice can be thought of as corresponding to holes of different sizes in a sieve, where nodes from high-density regions will easily sieve through to the local network. Therefore, regions of relatively high values in $w(U_i, -)$, say, over $S \subset [0, 1]$, should induce a "stretching" of the interval around S, and likewise, a "compression" of the interval outside of S.

Lemma 1. *As we know from the definition of a graphon, $f_{U_i}(x) = 1$ over $[0, 1]$. However,*

$$U_j^x \sim f_{U_j}(y \mid A_{ij} = 1, U_i = x) = \frac{w(x, y)}{\int_0^1 w(x, y)\,\mathrm{d}y}.$$

In the case of an SBM, we see that the multinomial class probabilities $(\alpha_1, \ldots, \alpha_Q)$ are weighted by the action of filtering through node i, resulting in a new set of multinomial class probabilities $(\pi_{Z_i,1}\alpha_1, \ldots, \pi_{Z_i,Q}\alpha_Q)/\bar{\pi}_{Z_i}$.

In fact, we may reformulate all of our results from the previous section in terms of this localized SBM with parameters $\pi' = \pi$ and $\alpha' = (\pi_{Z_i,1}\alpha_1, \ldots, \pi_{Z_e,Q}\alpha_Q)/\bar{\pi}_{Z_e}$. Indeed, $\check{\pi}_{Z_e s}$ can be related to $\bar{\pi}'_{Z_e s}$ in the following way:

Remark 2. *An expression for $\check{\pi}_{Z_e s}$ is*

$$\check{\pi}_{Z_e s} = \sum_{t \in [Q]} \pi'_{st}\alpha'_t = \bar{\pi}'_{Z_e s}.$$

The analysis in [5] can then be repeated with these values of π' and α'. This will yield results that are comparable to, but not exactly the same as what was derived in an analysis of the mutual friend counts [13]. The difference is that $n' \neq n$; that is, the size of the graph is now the size of the ego network, and not the size of the whole network. Since the size of the ego network is stochastic (as it equals the degree of the ego node), an analysis using π' and α' as the parameters of an SBM would essentially be conditioning on the size of the ego network being fixed at n'.

2.2 Representations of Local Graphons

In order to specify the node parametrization of the local graphon the canonical way, we need $U_j^x \sim \mathrm{Unif}(0, 1)$. As we know from basic probability theory, if given a random variable $X \sim F$, then a uniform random variable U can be transformed via $Q(U) \overset{d}{=} X$, where Q is called the *quantile function*. If the CDF is strictly monotone, we can invert it to get its quantile function, but if it has regions of zero derivative or jumps, we may use the generalized inverse [7] to get

the quantile function, instead. The generalized inverse of the CDF F is defined as

$$Q(p) \equiv \inf\{t : p \leq F(t)\}.$$

To summarize, we consider the *local graphon at* x, denoted by w_x, to mean the graphon obtained by filtering w through the slice $w(U_i = x, -)$.

Theorem 1. *The local graphon function can be represented as*

$$w_x(u, v) = w(Q_x(u), Q_x(v)).$$

We write the quantile function Q as Q_x with the subscript x since it will be useful to track the slice location $U_i = x$. And so, the probability of two nodes i and j connecting in a graph on n nodes drawn from the local graphon w_x, which we will denote by G_n^x with adjacency A_{ij}^x, is

$$A_{ij}^x | U_i^x = u, U_j^x = v \sim \text{Bern}(w_x(u, v) = w(Q_x(u), Q_x(v))).$$

A technical consideration when dealing with graphons is that any given graphon function w is merely a representation of an equivalence class of graphons, denoted $[w]$. For any two graphon functions w_1 and w_2 from $[w]$, there exist two *measure-preserving transformations*[1] ϕ_1 and ϕ_2, so that for all $u, v \in [0, 1]$,

$$w_1(\phi_1(u), \phi_1(v)) = w_2(u, v) \qquad w_1(u, v) = w_2(\phi_2(u), \phi_2(v)).$$

In other words, any two representations of the same graphon will be the same up to a kind of permutation of the domains.

A natural question to ask is whether two representations $w_1, w_2 \in [w]$ (related via a measure-preserving transformation $\phi : w_1(u, v) = w_2(\phi(u), \phi(v))$) with corresponding ego parameters x and $\phi(x)$ localize to two representations of a single local graphon. That is, are the localized versions of w_1 and w_2 still related via a measure-preserving transformation (possibly distinct from ϕ)?

Let Q_x be the quantile function of the distribution of U_j^x on the graphon function w_1, as defined above. Then,

$$\begin{aligned}(w_1)_x(u, v) &= w_1(Q_x(u), Q_x(v)) &(1)\\ &= w_2(\phi(Q_x(u)), \phi(Q_x(v))) &(2)\\ &\overset{?\exists\psi}{=} (w_2)_{\phi(x)}(\psi(u), \psi(v)), &(3)\end{aligned}$$

where

$$\phi(Q_x(u)) = \phi\left(\inf\left\{b : u \leq \frac{\int_0^b w_1(x, t)\,\mathrm{d}t}{\int_0^1 w_1(x, t)\,\mathrm{d}t}\right\}\right) = \phi\left(\inf\left\{b : u \leq \frac{\int_0^b w_2(\phi(x), \phi(t))\,\mathrm{d}t}{\int_0^1 w_2(\phi(x), \phi(t))\,\mathrm{d}t}\right\}\right).$$

If such a measure-preserving ψ exists, then w_2 localized to $\phi(x)$ belongs to the same graphon class induced by the tuple (w_1, x), which we will denote $[w_1, x]$.

[1] A measure-preserving transformation on a measure space (X, Σ, μ) is a map $f : X \to X$ satisfying $\mu(f^{-1}(A)) = \mu(A)$ for all $A \in \Sigma$.

Ideally, it shall always be the case that such a ψ will exist for any other localized representation of $[w]$. Indeed, we can write

$$(w_2)_{\phi(x)}(\psi(u), \psi(v)) = w_2(Q'_{\phi(x)}(\psi(u)), Q'_{\phi(x)}(\psi(v))),$$

where

$$Q'_{\phi(x)}(\psi(u)) \equiv \inf \left\{ b : \psi(u) \leq \frac{\int_0^b w_2(\phi(x), t)\,\mathrm{d}t}{\int_0^1 w_2(\phi(x), t)\,\mathrm{d}t} \right\}.$$

This yields a relation which defines ψ: for all $u, v \in [0, 1]$,

$$w_2(\phi(Q_x(u)), \phi(Q_x(v))) = w_2(Q'_{\phi(x)}(\psi(u)), Q'_{\phi(x)}(\psi(v))).$$

That this relation defines a valid measure-preserving map between these two graphon functions is established by the following theorem.

Theorem 2. *If ϕ is invertible and $w_1, w_2 > 0$ (so that $Q_{\phi(x)} \equiv F_x^{-1}$ and $Q'_x \equiv F'^{-1}_{\phi(x)}$), then $\psi \equiv F'_{\phi(x)} \circ \phi \circ F_x^{-1}$ is a measure-preserving map between $(w_1)_x$ and $(w_2)_{\phi(x)}$. Any two such local graphon functions are therefore representations of the same local graphon.*

To prove this, we will require the following lemma.

Lemma 2. *Let (X, Σ, μ) be a measure space with f integrable and $\phi : X \to X$ measure-preserving, i.e. $\mu(\phi^{-1}(S)) = \mu(S)\ \forall S \in \Sigma$. Then,*

$$\int_{\phi(S)} f\,\mathrm{d}\mu = \int_S f \circ \phi\,\mathrm{d}\mu.$$

Proof. Clearly $\forall f'$ integrable,

$$\int_X f'\,\mathrm{d}\mu = \int_X f' \circ \phi\,\mathrm{d}\mu.$$

Take $f'(x) \equiv f(x)1_{\phi(S)}(x)$. Then,

$$\begin{aligned}\int_{\phi(S)} f\,\mathrm{d}\mu &= \int_X f 1_{\phi(S)}\,\mathrm{d}\mu \\ &= \int_X (f \circ \phi)(1_{\phi(S)} \circ \phi)\,\mathrm{d}\mu.\end{aligned}$$

as ϕ is assumed to be injective, $\phi(x) \in \phi(S)$ implies that $x \in S$. So, $1_{\phi(S)} \circ \phi \equiv 1_S$. Hence,

$$\begin{aligned}\int_X (f \circ \phi)(1_{\phi(S)} \circ \phi)\,\mathrm{d}\mu &= \int_X (f \circ \phi) 1_S\,\mathrm{d}\mu \\ &= \int_S f \circ \phi\,\mathrm{d}\mu.\end{aligned}$$

□

Additional properties of measure-preserving maps on the unit inverval $[0,1]$ can be found in [6]. However, Lemma 2 is all we will need for the proof of Theorem 2.

Proof of Theorem 2. Note that we will use $w_1(x,-)$ and $w_2(\phi(x),-)$ to mean the cross-sections of the graphon, i.e. $w_1(x,-)(y) = w_1(x,y)$. Then, we know that

$$w_1(x,-) \equiv w_2(\phi(x),-) \circ \phi, \tag{4}$$

by the definition of ϕ. Define $\forall S \in \mathcal{B}[0,1]$

$$F_x(S) = \frac{\int_S w_1(x,-)\,\mathrm{d}\mu}{\int_0^1 w_1(x,-)\,\mathrm{d}\mu}$$
$$F'_{\phi(x)}(S) = \frac{\int_S w_2(\phi(x),-)\,\mathrm{d}\mu}{\int_0^1 w_2(\phi(x),-)\,\mathrm{d}\mu}.$$

Since ϕ is measure-preserving, we know (by integrating both sides of (4))

$$\int_0^1 w_1(x,-)\,\mathrm{d}\mu = \int_0^1 w_2(\phi(x),-)\circ\phi\,\mathrm{d}\mu = \int_0^1 w_2(\phi(x),-)\,\mathrm{d}\mu,$$

so the normalization factors cancel in $F'_{\phi(x)} \circ \phi \circ F_x^{-1}$.

Next, we will show that sampling a graphon from $(w_1)_x$ is equivalent to sampling a graphon from $(w_2)_{\phi(x)}$. Notice that if

$$F'_{\phi(x)} \circ \phi \circ F_x^{-1}(U) \overset{d}{=} U',$$

where $U, U' \sim \mathrm{Unif}(0,1)$, then we are done, as this establishes that $F'_{\phi(x)} \circ \phi \circ F_x^{-1}$ is a measure-preserving transformation. Rewriting,

$$\begin{aligned}(F'_{\phi(x)} \circ \phi \circ F_x^{-1})(U) &\overset{d}{=} U' \\ \iff F_x^{-1}(U) &\overset{d}{=} (F'_{\phi(x)} \circ \phi)^{-1}(U'),\end{aligned}$$

but this follows from observing that the LHS is an application of the inverse CDF to a uniform random variable. Ignoring the normalization factors, which we showed cancel, we see that $\forall a \in [0,1]$,

$$\begin{aligned}\mathbb{P}\left(F_x^{-1}(U) \in [0,a)\right) &= \int_{[0,a)} w_1(x,-)\,\mathrm{d}\mu \\ &= \int_{[0,a)} w_2(\phi(x),-)\circ\phi\,\mathrm{d}\mu.\end{aligned}$$

By Lemma 2,

$$\begin{aligned}\int_{[0,a)} w_2(\phi(x),-)\circ\phi\,\mathrm{d}\mu &= \int_{\phi([0,a))} w_2(\phi(x),-)\,\mathrm{d}\mu \\ &= \mathbb{P}\left(F'^{-1}_{\phi(x)}(U') \in \phi([0,a))\right) \\ &= \mathbb{P}\left((F'_{\phi(x)} \circ \phi)^{-1}(U') \in [0,a)\right).\end{aligned}$$

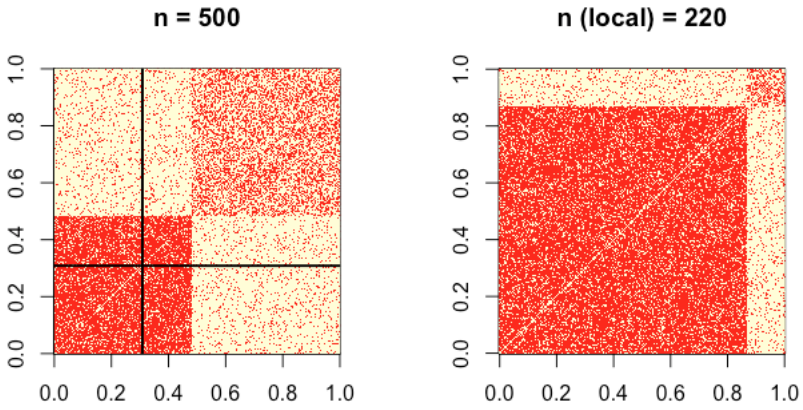

Figure 1: Realization of the ambient graphon (left), actually an SBM with parameters $\pi_{11} = 0.8$, $\pi_{22} = 0.4$, $\pi_{12} = 0.1$, and $\alpha_1 = \alpha_2 = 0.5$. The black lines correspond to the slice $w(U_1 \approx 0.3, -)$ at the parameter of the first-index node in the sampled graph. Visually, the membership probabilities for the localized SBM (right) are approximately $\alpha'_1 = 0.87$ and $\alpha'_2 = 0.13$, whereas theoretically we expect $\alpha'_1 = \pi_{11}\alpha_1/(\pi_{11}\alpha_1 + \pi_{12}\alpha_2) = 0.\overline{8}$ and $\alpha'_1 = 0.\overline{1}$.

So, $F'_{\phi(x)} \circ \phi$ is the same CDF as F_x. □

2.3 Some Example Realizations of Localized Graphons

Using the same process as described above for generating a local network, we sample local networks from some simple examples for the ambient graphon $[w]$. The results of these small experiments[2] are displayed in Figures 1 and 2 as empirical graphons. Figure 3 shows the KDE-smoothed versions of these examples. These plots were generated as follows.

1. Fix an ambient graphon representation w.
2. Sample a graph from w with n nodes, with adjacency matrix A.
3. Plot A as an empirical graphon, with the nodes sorted by the true values of U_i.
4. Let A^x be the principal submatrix of A obtained by keeping only the rows/columns indexed by j where $A_{1j} = 1$.
5. Plot A^x as an empirical graphon, again with the nodes sorted by U_i.

[2]The R code [19] used to generate these figures can be downloaded at `https://www.stat.cmu.edu/~nicolask/share/localgraphon`.

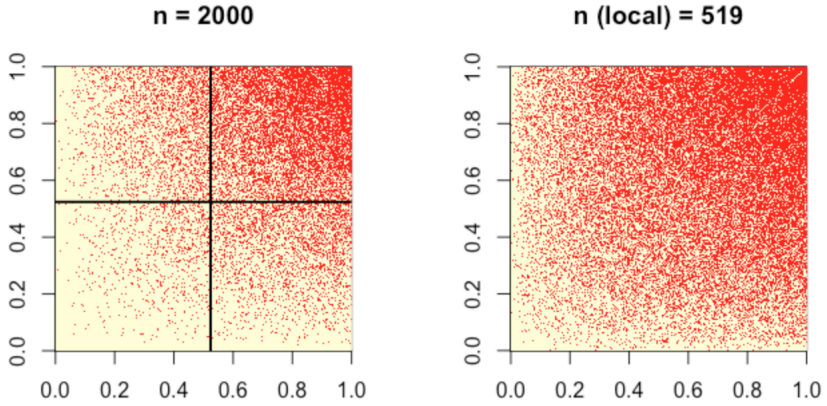

Figure 2: Realization of the ambient graphon (left), given by $w(u, v) = uv$. The black lines correspond to the slice $w(U_1 \approx 0.5, -)$ at the parameter of the first-index node in the sampled graph. The local graphon (right) induced by this slice has a greater density of points in the lower and left-hand edges, since the sparse regions are under-represented in the local graphon.

3 Snowball Sampling

As a natural extension of the case of the graphon induced by the ego network, we can describe the local graphon induced by more than one ego, or seed, node. Let W_0 denote the set of seed nodes, called the *seed set*, from which *waves* $W_1, W_2, \ldots$ of snowball samples will grow. We (recursively) define a wave via

$$W_k \equiv \left\{ i \in V(G) : \sum_{j \in W_{k-1}} A_{ij} > 0, \sum_{j \in \bigcup_{\ell < k-1} W_\ell} A_{ij} = 0 \right\}.$$

In other words, any node in the kth wave must be linked to at least one node in the $(k-1)$th wave, and it must not be linked to any node in a previous wave.

Theorem 3. *In general, if there is a set of nodes W^- that node j should not connect to, and a set of nodes W^+ that node j should have at least one connection to, then:*

$$\mathbb{P}\left(U_j = y \,\middle|\, \sum_{i \in W^-} A_{ij} = 0, \sum_{i \in W^+} A_{ij} > 0, \{U_i\}_{i \in W^- \cup W^+} \right) = \frac{(1 - \prod_{i \in W^+}(1 - w(u_i, y))) \prod_{e \in W^-}(1 - w(u_e, y))}{\int_0^1 (1 - \prod_{i \in W^+}(1 - w(u_i, y))) \prod_{e \in W^-}(1 - w(u_e, y)) \, \mathrm{d}y}.$$

Corollary 2. *The local graphon induced by multiple slices is described by*

$$U_j^{W_0} \sim f_{U_j}\left(y \,\middle|\, \sum_{e \in W_0} A_{ej} > 0, \{U_e\}_{e \in W_0} \right) = \frac{1 - \prod_{e \in W_0}(1 - w(u_e, y))}{\int_0^1 1 - \prod_{e \in W_0}(1 - w(u_e, y)) \, \mathrm{d}y}.$$

Clearly, when $W_0 = \{i\}$, this reduces to Lemma 1.

Furthermore, the size of the first wave is:

$$|W_1| \mid W_0, \{U_e\}_{e \in W_0} \sim \text{Bin}\left(n - |w_0|, \int_0^1 1 - \prod_{e \in w_0} (1 - w(u_e, y)) \, \mathrm{d}y\right).$$

If $|W_0| = 1$, this is just the size of the ego network, as in Corollary 1. Similarly, the size of second wave is described by:

$$|W_2| \mid W_0, W_1, \{U_e\}_{e \in W_0}, \{U_i\}_{i \in W_1} \sim \text{Bin}\left(n - |w_0| - |w_1|, \int_0^1 \left(\prod_{e \in w_0} (1 - w(u_e, y))\right)\left(1 - \prod_{i \in w_1} (1 - w(u_i, y))\right) \, \mathrm{d}y\right).$$

For each node not in the seed set or in the first wave, the probability of it entering the second wave is equal to the probability of it not connecting to the seed set, and then connecting to at least one node in the first wave. One can easily extend this to find the size distribution of the kth wave.

Lemma 3. *In general, the size of the kth wave is distributed as*

$$|W_k| \mid \{W_\ell\}_{\ell < k}, \{U_i\}_{i \in \bigcup_{\ell < k} W_\ell} \sim \text{Bin}\left(n - \sum_{i=0}^{k-1} |w_i|, \int_0^1 \left(\prod_{i \in \bigcup_{j=0}^{k-2} w_j} (1 - w(u_i, y))\right)\left(1 - \prod_{i \in w_{k-1}} (1 - w(u_i, y))\right) \, \mathrm{d}y\right).$$

Note that the subgraph induced by $\bigcup_{\ell \leq k} W_\ell$ is not a graphon, since the existence of some edge to W_{k-1} for every node in W_k is guaranteed via this construction. However, the subgraph induced by any given wave W_k is a graphon random graph.

4 Unbiased Graphon Estimation from Snowball Samples

Definitions for snowball sampling are provided by [9, 10, 14]. The particular design we consider for now is a one-wave snowball sample starting from a single ego node. This is just a different way of stating the scenario posed in the previous sections. The sample will consist of all of the alters of the one ego node; we do not include the ego node in this sample.

Some of the network parameters will be impossible to estimate from just one of these samples. Even asymptotically (as the size of the whole network grows to infinity), any SBM parameters associated to communities for which the ego's community has zero probability of connecting to will be fully unobserved, and therefore cannot be estimated. For graphons, any areas that coincide with regions of zero density along the ego's slice will be impossible to estimate. A relative understanding of the model, one limited to those regions of positive density, is still possible. However, to make the following analysis simpler, we assume that all entries of π are strictly positive; likewise, we assume $w(u, v) > 0$ for all $u, v \in [0, 1]$. This has an additional implication for the graphon case: $F_x(u)$ is strictly increasing and therefore its associated quantile function is simply its inverse.

For the SBM, recall:

$$\begin{cases} \alpha'_1 = \frac{\alpha_1 \pi_{Z_e,1}}{\bar{\pi}_{Z_e}} \\ \vdots \\ \alpha'_Q = \frac{\alpha_Q \pi_{Z_e,Q}}{\bar{\pi}_{Z_e}} \end{cases} \overset{\text{Invert}}{\Longrightarrow} \begin{cases} \alpha_1 = \frac{\alpha'_1 \bar{\pi}_{Z_e}}{\pi_{Z_e,1}} \\ \vdots \\ \alpha_Q = \frac{\alpha'_Q \bar{\pi}_{Z_e}}{\pi_{Z_e,Q}}. \end{cases}$$

A standard SBM estimation algorithm will provide estimates of π and α' given a sampled ego network. However, Z_e and $\bar{\pi}_{Z_e}$ are not known *a priori*. By estimating Z_e, an estimate for $\bar{\pi}_{Z_e}$ can be obtained via $\sum_q \alpha_q = 1$. In fact, using only the mutual friend counts, Z_e is non-identifiable in some cases. However, as long as $\check{\pi}_{Z_e Z_e} > \check{\pi}_{Z_e s}$ for all blocks s, Z_e can be estimated. This holds under the conditions discussed in Section 3.1 of [13].

This means we are only missing an estimate of $\bar{\pi}_{Z_e}$. Since it is only a normalization factor, we could estimate it via

$$\hat{\bar{\pi}}_{Z_e} = \sum_{q \in [Q]} \pi_{\hat{Z}_e q} \hat{\alpha}_q.$$

However, this assumes that nodes from every community $q = 1, \ldots, Q$ is observed in the ego network. Instead, we can try to estimate the *observed* subset of the SBM.

The problem carries over naturally to the more general graphon case. For the graphon, one can just reverse the defining equation for the local graphon,

$$w_x(u, v) = w(Q_x(u), Q_x(v))$$

to obtain

$$w(u', v') = w_x(F_x(u'), F_x(v')).$$

The unknown here is the parameter of the ego node, $U_i = x$. Therefore, one may first run a graphon estimation algorithm (e.g. kernel-density smoothing of the empirical graphon) to obtain an estimate for w_x; all that remains is to estimate U_i. U_i cannot be identified from any one realization of the graphon without making certain assumptions about w, such as the graphon being canonical, as in [4].

The following theorem demonstrates that given an estimate of the local position of the ego node's location, $u_x = F_x(x)$, it is possible to undo the localization distortion in w_x to obtain an estimate of w.

Theorem 4. *Let u_x be an estimate of $F_x(x)$, that is, the location parameter in the local graphon space that corresponds to x, the location of the ego node, in the ambient graphon space. Furthermore, let $w > 0$, and in addition, that $w(x, -)$ is continuous. Then,*

$$w(u, v) = w_x\left(F_x(u), F_x(v)\right),$$

where

$$F_x(y) = \inf \left\{ \alpha : y \leq \frac{\int_0^\alpha \frac{1}{w_x(u_x, t)} \, \mathrm{d}t}{\int_0^1 \frac{1}{w_x(u_x, t)} \, \mathrm{d}t} \right\}.$$

Proof. Since $w_x(u_x, v) = w(x, F_x^{-1}(v))$, and

$$F_x(y) = \int_0^y \frac{w(x,t)}{\int_0^1 w(x,t)\,\mathrm{d}t}\,\mathrm{d}t,$$

we have

$$w_x(u_x, v) = \left(\int_0^1 F_x'(t)\,\mathrm{d}t\right) \times F_x' \circ F_x^{-1}.$$

By the inverse function theorem,

$$F_x' \circ F_x^{-1} = \frac{1}{(F_x^{-1})'},$$

so

$$F_x^{-1}(y) = \int_0^y \frac{\int_0^1 w(x,t)\,\mathrm{d}t}{w_x(u_x, \beta)}\,\mathrm{d}\beta.$$

To estimate the numerator, we can use the fact that $F_x^{-1}(1) = 1$ to obtain

$$\int_0^1 w(x,t)\,\mathrm{d}t = \frac{1}{\int_0^1 \frac{1}{w_x(u_x,t)}\,\mathrm{d}t}.$$

This proves the result. □

5 Conclusions

A better understanding of the sampling bias induced by sampling edges rather than nodes could lead to more efficient estimation procedures for large social networks. In particular, computational limitations involved with studying networks with billions of nodes may necessitate subsampling. This work establishes some of the basic properties of "local" subsamples of large networks which may be modeled as graphon random graphs.

Theorem 2 provides a partial answer to the question: "Are any two representations of the same local graphon in the same equivalence class as graphons?" In the case where ϕ is invertible the answer is yes, but in general two graphon functions w_1 and w_2 representing the same graphon will be related via two measure-preserving maps ϕ_1 and ϕ_2 which are not necessarily inverses (nor even invertible). In our proof, we leverage knowledge of the CDFs induced by the graphon slices to equate two distributions; it is possible that slight modification of the same argument will establish the result without needing to invert ϕ. In addition, we currently impose the restriction that $w_1, w_2 > 0$ so that the CDFs induced by the graphon slices are strictly increasing (and therefore invertible). However, it should be possible to remove this assumption; the regions where $w_1(x, -) = 0$ should be the regions where $w_2(\phi(x), -) \circ \phi = 0$, so one could invert on a restriction of the domain and make a separate argument for the flat regions of the CDFs.

Acknowledgments

We thank the Stat-Networks group at Carnegie Mellon and Peter Elliott for many valuable discussions throughout the development of this work. This work was partially supported by AFOSR grant #FA9550-14-1-0141.

References

[1] Edoardo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

[2] Patrick Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, third edition, 1995.

[3] Diana Cai, Nathanael Ackerman, and Cameron Freer. An iterative step-function estimator for graphons. *arXiv preprint arXiv:1412.2129*, 2014.

[4] Stanley H Chan and Edoardo M Airoldi. A consistent histogram estimator for exchangeable graph models. In *ICML*, pages 208–216, 2014.

[5] Antoine Channarond, Jean-Jacques Daudin, and Stéphane Robin. Classification and estimation in the stochastic blockmodel based on the empirical degrees. *Electronic Journal of Statistics*, 6:2574–2601, 2012.

[6] Behrouz Emamizadeh et al. The distribution function and measure preserving maps. *Real Analysis Exchange*, 36(1):161–168, 2010.

[7] Paul Embrechts and Marius Hofert. A note on generalized inverses. *Mathematical Methods of Operations Research*, 77(3):423–432, 2013.

[8] Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, Dec 2015.

[9] Leo A Goodman. Snowball sampling. *The annals of mathematical statistics*, pages 148–170, 1961.

[10] Mark S Handcock and Krista J Gile. Comment: On the concept of snowball sampling. *Sociological Methodology*, 41(1):367–371, 2011.

[11] Douglas D Heckathorn. Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems*, 44(2):174–199, 1997.

[12] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, Jun 1983.

[13] Nicolas Kim and Alessandro Rinaldo. Community detection on ego networks via mutual friend counts. Preprint, 2017.

[14] Eric D. Kolaczyk. Statistical analysis of network data. *Springer Series in Statistics*, 2009.

[15] Pavel N Krivitsky and Martina Morris. Inference for social network models from egocentrically-sampled data, with application to understanding persistent racial disparities in hiv prevalence in the us. 2015.

[16] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Physical Review E*, 73(1):016102, 2006.

[17] László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

[18] László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[20] Tom A.B. Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, Jan 1997.

[21] Kyle Shane Vincent. *Strategies for estimating the size and distribution of hard-to-reach populations with adaptive sampling*. PhD thesis, Simon Fraser University, 2012.

[22] Cyprian Wejnert. Social network analysis with respondent-driven sampling data: A study of racial integration on campus. *Social Networks*, 32(2):112–124, 2010.

[23] Patrick J Wolfe and Sofia C Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

A Appendix

A.1 Conditions on π and α

To make the problem simpler, we assume that the block matrix π is of the form

$$\pi = \begin{bmatrix} w & b & \dots & b \\ b & w & \dots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \dots & w \end{bmatrix}$$

so that all of the within-block probabilities are w, and all of the between-block probabilities are b (similar results can be found without this assumption, but to keep this document short we omit them). This means that

$$\begin{aligned} \check{\pi}_{Z_e s} &= \frac{1}{\bar{\pi}_{Z_e}} \left[\left(\sum_{t \in [Q] \setminus \{Z_e, s\}} \pi_{Z_e t} \pi_{st} \alpha_t \right) + \pi_{Z_e s} (\pi_{Z_e Z_e} \alpha_{Z_e} + \pi_{ss} \alpha_s) \right] \\ &= \frac{1}{\bar{\pi}_{Z_e}} \left[\left(\sum_{t \in [Q] \setminus \{Z_e, s\}} b^2 \alpha_t \right) + wb(\alpha_{Z_e} + \alpha_s) \right] \end{aligned}$$

and

$$\begin{aligned} \check{\pi}_{Z_e Z_e} &= \frac{1}{\bar{\pi}_{Z_e}} \left[\left(\sum_{t \in [Q] \setminus \{Z_e\}} \pi_{Z_e t}^2 \alpha_t \right) \pi_{Z_e Z_e}^2 \alpha_{Z_e} \right] \\ &= \frac{1}{\bar{\pi}_{Z_e}} \left[\left(\sum_{t \in [Q] \setminus \{Z_e\}} b^2 \alpha_t \right) + w^2 \alpha_{Z_e} \right]. \end{aligned}$$

Then, the necessary and sufficient conditions for the ego's community to have the highest expected mutual friend count are

$$\begin{cases} \check{\pi}_{Z_e Z_e} - \check{\pi}_{Z_e 1} > 0 \\ \check{\pi}_{Z_e Z_e} - \check{\pi}_{Z_e 2} > 0 \\ \vdots \\ \check{\pi}_{Z_e Z_e} - \check{\pi}_{Z_e (Z_e - 1)} > 0 \\ \check{\pi}_{Z_e Z_e} - \check{\pi}_{Z_e (Z_e + 1)} > 0 \\ \vdots \\ \check{\pi}_{Z_e Z_e} - \check{\pi}_{Z_e Q} > 0 \end{cases} \iff \begin{cases} (w-b)(w\alpha_{Z_e} - b\alpha_1) > 0 \\ (w-b)(w\alpha_{Z_e} - b\alpha_2) > 0 \\ \vdots \\ (w-b)(w\alpha_{Z_e} - b\alpha_{Z_e - 1}) > 0 \\ (w-b)(w\alpha_{Z_e} - b\alpha_{Z_e + 1}) > 0 \\ \vdots \\ (w-b)(w\alpha_{Z_e} - b\alpha_Q) > 0 \end{cases} \iff \begin{cases} \alpha_{Z_e} > \frac{b}{w} \alpha_1 \\ \alpha_{Z_e} > \frac{b}{w} \alpha_2 \\ \vdots \\ \alpha_{Z_e} > \frac{b}{w} \alpha_{Z_e - 1} \\ \alpha_{Z_e} > \frac{b}{w} \alpha_{Z_e + 1} \\ \vdots \\ \alpha_{Z_e} > \frac{b}{w} \alpha_Q. \end{cases}$$

Assuming an assortative SBM, so that $w > b$, the ego community can be quite small in the overall network. Generally, w is taken such that $w \gg b$, so that the ego community will generally be identifiable from the mutual friend counts. Note that even if $\alpha'_{Z_e} = \alpha'_s$ for some $s \neq Z_e$, the mutual friend counts are still expected to be higher for the ego's community, since $\pi_{Z_e Z_e} > \pi_{Z_e s}$. This property endows the `Mutual-Friends` algorithm with a kind of robustness, as long as the whole network can be adequately modeled with an SBM.

A.2 Proof of Lemma 1

Proof. Intuitively,

$$f_{U_j}(y|A_{ij}=1,U_i=x) = \frac{\mathbb{P}\left(A_{ij}=1|U_i=x,U_j=y\right) f_{U_j}(y|U_i=x)}{\mathbb{P}\left(A_{ij}=1,U_i=x\right)}$$
$$= \frac{w(x,y)}{\int_0^1 w(x,y)\,\mathrm{d}y}.$$

Note the conditioning on a zero-measure set, namely that U_i takes on a particular fixed value. Since the unit inverval is separable and complete under the usual topology and metric, it admits regular conditional probabilities [2]. In particular, if

$$\mathbb{P}\left(A_{ij}=1,U_i=x\right) = \int_0^1 w(x,y)\,\mathrm{d}y > 0$$

for λ-almost every $x \in [0,1]$, then $f_{U_j}(y|A_{ij}=1,U_i=x)$ is a valid probability density for almost every x. This means we can take the probability measure we are integrating with respect to to be the appropriate regular version. □

A.3 Defining the Local Network Generative Process

To make rigorous the notion of a local graphon, we define the process which generates this object.

1. Given a graphon function w representing the graphon $[w]$, sample a sequence of graphs $G_n \sim w$ for $n \in [N]$, where $G_n = G_{n-1} + g_n$, where g_n is a one-node graph sampled from w. That is, we add one node at each time step n to the existing graph, and the edges to g_n are determined via the graphon function. G_0 is defined to be the empty graph, $G_0 = (V,E) = (\emptyset,\emptyset)$.
2. WLOG (due to exchangeability), consider the first node, i.e. the only node in G_1. Condition on its parameter $U_1 = x$, for some fixed $x \in [0,1]$.
3. The expected number of nodes in G_n connected to the first node (henceforth the *ego node*) is therefore

$$\mathbb{E}\left[\sum_{j\neq 1} A_{1j} \,\middle|\, U_1 = x\right] = (n-1)\int_0^1 w(x,y)\,\mathrm{d}y = (n-1)d_w(x) \to \infty,$$

   if $d_w(x) > 0$.
4. Hence, the graphs in the sequence of subgraphs G_n^x induced by the vertex set $\{j \in V(G_n) : A_{1j} = 1\}$ almost surely grow unboundedly in size, and the convergence of homomorphism densities of G_n should induce convergence of G_n^x (although this remains to be rigorously established).

We can therefore associate to a given representation w of $[w]$ a local graphon induced by nodes filtered through correspondence with an ego node having parameter x, where x is relative to the particular representation.

A.4 Smoothed Local Graphon Estimates

Figures 1 and 2 displayed the empirical graphons as zero-one estimates. In Figure 3 we apply KDE smoothing to obtain possibly clearer depictions of the true local graphon functions.

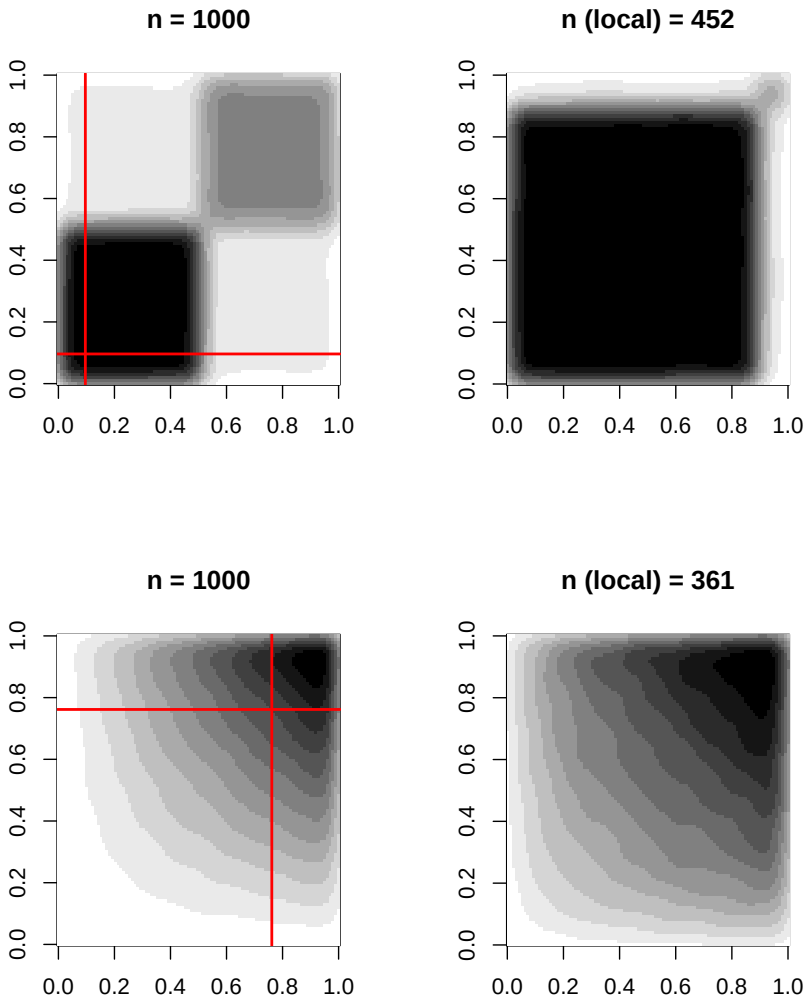

Figure 3: These smoothed empirical graphons are more accurate depictions of the true graphons. The slope graphon $w(u, v) = uv$ is more clearly skewed by the localization procedure than in the zero-one depiction from Figure 2.