

Poisson approximation of subgraph counts in stochastic block models and a graphon model

Matthew Coulson*, Robert E. Gaunt[†] and Gesine Reinert[†]

September 28, 2015

Abstract

Small subgraph counts can be used as summary statistics for large random graphs. We use the Stein-Chen method to derive Poisson approximations for the distribution of the number of subgraphs in the stochastic block model which are isomorphic to some fixed graph. We also obtain Poisson approximations for subgraph counts in a graphon-type generalisation of the model in which the edge probabilities are (possibly dependent) random variables supported on a subset of $[0, 1]$. Our results apply when the fixed graph is a member of the class of strictly balanced graphs.

Keywords: Graphon model, stochastic block model, Erdős-Rényi Mixture model, subgraph counts, Poisson approximation, Stein-Chen method

AMS 2010 Subject Classification: 90B15, 62E17, 60F05, 05C80

1 Introduction

Small subgraph counts can be used as summary statistics for large random graphs; indeed in some graph models they appear as sufficient statistics, see [12]. Moreover, many networks are conjectured to have over- or under-represented motifs (small subgraphs), see for example [19]. Statistics based

*The Queen's College, University of Oxford, High Street, OXFORD, OX1 4AW, UK

[†]Department of Statistics, University of Oxford, 1 South Parks Road, OXFORD OX1 3TG, UK

on small subgraph counts can be used to compare networks, as in [3, 24]. To determine which small subgraphs are unusual, assessing the distribution of such motifs is key. While [22] gives the mean and variance for some common random graph models, [22] does not derive a distributional approximation.

In this paper, we address the issue of such a distributional approximation for a large class of models which include stochastic block models and graphon models but also models with random edge probabilities, provided that the edge probabilities display some local dependence, which will be made clearer in Section 3.

The stochastic block model (SBM) was introduced originally for directed graphs by [13] and generalised to other graphs by [20]; it is also called Erdős-Rényi Mixture Model in [10], and in theoretical computer science it is called the Planted Partition Model [9]. It has a wide range of applications, see for example [1, 6, 10, 14, 21], and [17] for a recent survey. The model is defined as follows. Consider an undirected random graph on n vertices, with no self-loops or multiple edges, in which the vertices are spread among Q hidden classes with respective proportion vector $f = (f_1, \dots, f_Q)$. The class label of a vertex is drawn from a multinomial distribution $\mathcal{M}(1, f)$. Edges $Y_{i,j}$ are independent conditionally to the class of the vertices, and the edge probability depends only on the classes of the vertices:

$$\mathbb{P}(Y_{i,j} = 1 \mid i \in a, j \in b) = \pi_{a,b}.$$

We shall denote this model by $SBM(n, \pi, f)$. If $\pi_{a,b} = p$ for all a and b , the SBM reduces to the classical Erdős-Rényi random graph model, which we denote by $\mathcal{G}(n, p)$. In this paper, it is assumed that π and f are known, and that $f_1, \dots, f_Q > 0$; for estimating these quantities see, for example, [1, 15, 21].

For a fixed graph G , it is known (see, for example, [5], Theorem 5.B) that the distribution of the number copies of G in the $\mathcal{G}(n, p)$ model is well approximated by an appropriate Poisson distribution if G is a member of the class of strictly balanced graphs (defined below) as long as p is not too large.

In this paper, we consider a generalisation of the Poisson approximation to the stochastic block model. We consider both the cases that the edge probabilities $\pi_{a,b}$ are constant and that they are themselves random variables supported on a subset of $[0, 1]$.

When the edge probabilities are themselves random variables then we assume that they are only locally dependent, in the sense that each edge

has a relatively small number of other edges so that their edge probabilities are not independent. As an example the vertices may have some exogeneous characteristics such as geographical location which influence the probability of an edge to exist, but only locally.

The latter case is related to graphon models, where edge probabilities only depend on those edge probabilities where the edges share a vertex. A graphon is represented by a measurable function $h : [0, 1]^2 \rightarrow [0, 1]$. A graphon model constructs a random graph on n vertices by assigning independent $U[0, 1]$ variables to each vertex. Conditional on these uniform random variables, all edges are independent, and the probability of an edge between vertices u and v is given by $h(U_u, U_v)$. These graphs appear as limits of exchangeable graphs; see, for example, [2, 11, 16]. They are a special case of inhomogeneous random graph models as considered in [7].

Setting the scene for counting copies of graphs G , let K_n be the complete graph with n vertices and $\binom{n}{2}$ edges. Let $G \subset K_n$ be a fixed graph with $v(G)$ vertices and $e(G)$ edges; let $V(G)$ denote the vertex set and $E(G)$ its edge set. To avoid trivialities, we assume that $e(G) > 1$ and that G has no isolated vertices. We shall be particularly interested in the case that G is a member of the class of strictly balanced graphs, which we now define according to [5]. Let

$$d(G) = \frac{e(G)}{v(G)}.$$

Then the graph G is said to be *strictly balanced* if $d(H) < d(G)$ for all subgraphs $H \subsetneq G$.

Let Γ denote the set of $v(G)$ -tuples of elements from $\{1, \dots, n\}$. Then, $\alpha \in \Gamma$ is a possible position for the subgraph G , of which there are $\binom{n}{v(G)}$ such positions. To account for re-labelling of vertices, let $R(G)$ denote the set of all subgraphs of K_n which are isomorphic to G , and let $\rho(G) = |R(G)|$. Then

$$\rho(G) = \frac{(v(G))!}{a(G)}, \quad (1.1)$$

where $a(G)$ is the number of elements in the automorphism group of G .

Now, for $\alpha \in \Gamma$ and $G' \in R(G)$, let $X_\alpha(G')$ be the indicator random variable for the occurrence, at α , of a subgraph G' which is isomorphic to G . We shall let W denote the total number of copies of G in the random graph,

$$W = \sum_{\alpha \in \Gamma} \sum_{G' \in R(G)} X_\alpha(G'). \quad (1.2)$$

Here, copies are counted as opposed to induced copies where not only all edges of the graph have to appear, but also no edge which is not in the graph is allowed to appear in the copy. For example, the complete graph K_n , $n \geq 3$, contains $(n-1)!/2$ copies, but no induced copy, of an n -cycle.

In the stochastic block model $SBM(n, \pi, f)$, the conditional occurrence probability of the subgraph G on $\alpha = (i_1, \dots, i_{v(G)})$ given the class of each vertex is

$$\mathbb{P}(X_\alpha(G) = 1 \mid i_1 \in c_1, \dots, i_{v(G)} \in c_{v(G)}) = \prod_{1 \leq u < v \leq v(G): (u,v) \in E(G)} \pi_{c_u, c_v}.$$

The occurrence probability of G is then

$$\mu(G) = \mathbb{E}X_\alpha(G) = \sum_{c_1, c_2, \dots, c_{v(G)}=1}^Q f_{c_1} f_{c_2} \cdots f_{c_{v(G)}} \prod_{1 \leq u < v \leq v(G): (u,v) \in E(G)} \pi_{c_u, c_v}, \quad (1.3)$$

and

$$\lambda := \mathbb{E}W = \binom{n}{v(G)} \rho(G) \mu(G). \quad (1.4)$$

In this paper, we use the Stein-Chen method for Poisson approximation, introduced by [8], to assess the distributional distance between $\mathcal{L}(W)$ and the $Po(\lambda)$ distribution when the fixed graph G is a member of the class of strictly balanced graphs. This discrepancy is measured using the total variation distance, which for non-negative, integer-valued random variables U and V is given by

$$d_{TV}(\mathcal{L}(U), \mathcal{L}(V)) = \sup_{A \subseteq \mathbb{Z}^+} |\mathbb{P}(U \in A) - \mathbb{P}(V \in A)|.$$

In deriving bounds on the total variation distance, we exploit the local dependence structure of the indicators $X_\alpha(G)$. To this end, for each $\alpha \in \Gamma$, we introduce a set A_α which can be viewed as a dependency neighbourhood of α . When the edge probabilities are constants, we can take

$$A_\alpha = \{\beta \in \Gamma : |\alpha \cap \beta| \geq 1\}.$$

Here, A_α is a dependency neighbourhood of α in the sense that if $|\alpha \cap \beta| = 0$, then $X_\alpha(G)$ and $X_\beta(G')$ are independent for any $G, G' \in \mathcal{R}(G)$. In Section 3,

the edge probabilities are random variables, in which case finding a suitable dependency neighbourhood A_α is more involved. With

$$\eta_\alpha(G) = \sum_{\beta \in A_\alpha} \sum_{G' \in R(G)} X_\beta(G')$$

and

$$\theta_\alpha(G) = X_\alpha(G)(\eta_\alpha(G) - X_\alpha(G)), \quad (1.5)$$

a simple corollary of Theorem 1 in [4], or of Theorem 1.A in [5] is that

$$d_{TV}(\mathcal{L}(W), Po(\lambda)) \leq \lambda^{-1}(1 - e^{-\lambda}) \sum_{\alpha \in \Gamma} \sum_{G' \in R(G)} \{\mathbb{E}X_\alpha(G')\mathbb{E}\eta_\alpha(G') + \mathbb{E}\theta_\alpha(G')\}. \quad (1.6)$$

Thus bounding the total variation distance between the distribution of the subgraph counts in the SBM and the $Po(\lambda)$ distribution reduces to bounding the expectations on the right-hand side of (1.6). We shall prove our Poisson approximations for subgraph counts (Theorems 2.1, 3.1 and 4.1) using this approach. The Poisson approximation results of these theorems are valid when the fixed graph G is strictly balanced and the edge probabilities $\pi_{a,b}$ are not too large. These theorems generalise Theorem 5.B of [5], which asserts that a Poisson approximation is valid in the $\mathcal{G}(n, p)$ model under the same conditions.

The Poisson approximation is valid under these conditions in the SBM for exactly the same reason as it is in the $\mathcal{G}(n, p)$ model: if G is strictly balanced and the $\pi_{a,b}$ are not too large, with high probability the copies of G are vertex disjoint and the $X_\alpha(G)$ are close to being independent. Thus, W is the sum of a large number of almost independent indicators with small means, and a Poisson approximation is valid. In the $\mathcal{G}(n, p)$ model, the Poisson approximation breaks down if G is not strictly balanced [23], although Compound Poisson approximations may still be valid for certain classes of subgraphs; see [25]. For this reason, we restrict our attention to strictly balanced graphs.

The rest of the paper is organised as follows. In Section 2, we use the Stein-Chen method to derive a Poisson approximation for the number of subgraphs in the SBM which are isomorphic to some fixed graph from the class of strictly balanced graphs. In Section 3, we consider a generalisation of this problem in which the edge probabilities are now (possibly locally dependent) random variables supported on a subset of $[0, 1]$. Again, we derive a Poisson approximation for the number of copies of a fixed subgraph

in this model. Section 4 gives a Poisson approximation of small graph counts in the graphon model.

2 Poisson approximation of subgraph counts in the stochastic block model

In this section, we obtain a Poisson approximation for the number of subgraphs in the SBM which are isomorphic to a fixed graph from the class of strictly balanced graphs. Before stating this result, we introduce some notation. Let

$$\alpha(G) = \min_H \frac{e(G) - e(H)}{v(G) - v(H)} \quad (2.1)$$

and

$$\gamma(G) = \min_H (d(G)v(H) - e(H)) = \min_H v(H) \cdot (d(G) - d(H)), \quad (2.2)$$

where the minima are taken over all non-empty subgraphs $H \subsetneq G$ without isolated vertices. It is worth noting that the graph G is strictly balanced if $\gamma(G) > 0$ or $\alpha(G) > d(G)$; see [5]. Also, let

$$\pi^* = \max_{1 \leq a < b \leq Q} \pi_{a,b} \quad (2.3)$$

denote the maximum edge probability.

Theorem 2.1. *Suppose G is a strictly balanced graph. Then, with the notation (1.3), (1.1), (2.1), (2.2) and (2.3),*

$$\begin{aligned} d_{TV}(\mathcal{L}(W), Po(\lambda)) &\leq (1 - e^{-\lambda}) \rho(G) \left\{ 2 \frac{v(G)^2}{v(G)!} n^{v(G)-1} (\pi^*)^{e(G)} + \pi^* \right. \\ &\quad \left. + \sum_{s=2}^{v(G)-1} \binom{v(G)}{s} \frac{n^{v(G)-s} (\pi^*)^{\kappa(G,s)}}{(v(G)-s)!} \right\}, \end{aligned} \quad (2.4)$$

where

$$\kappa(G, s) = \max(e(G) - sd(G) + \gamma(G), (v(G) - s)\alpha(G)). \quad (2.5)$$

Proof. We establish our bound by bounding the right-hand side of inequality (1.6), starting with $\sum_{\alpha \in \Gamma} \sum_{G' \in R(G)} \mathbb{E}X_\alpha(G') \mathbb{E}\eta_\alpha(G')$. For the dependence set $A_\alpha = \{\beta \in \Gamma : |\alpha \cap \beta| \geq 1\}$,

$$|A_\alpha| \leq v(G) \binom{n}{v(G)-1} \leq \frac{v(G)^2}{v(G)!} n^{v(G)-1}. \quad (2.6)$$

It is now clear from (1.3) and (2.6) that

$$\begin{aligned} \mathbb{E}\eta_\alpha(G) &= \sum_{\beta \in A_\alpha} \sum_{G' \in R(G)} \mathbb{E}X_\beta(G') = |A_\alpha| |R(G)| \mu(G) \\ &\leq \frac{\rho(G) v(G)^2}{v(G)!} n^{v(G)-1} \mu(G). \end{aligned} \quad (2.7)$$

The more involved part of the proof, where the assumption of strictly balancedness comes into play, is to bound the expectation $\mathbb{E}\theta_\alpha(G)$ from (1.5). When α and β have considerable overlap, then $\mathbb{E}X_\alpha(G)X_\beta(G')$ may be large compared to $\mathbb{E}X_\alpha(G)$ - but there are not many β 's which have considerable overlap with α . To take account of the overlap, we partition A_α into sets $\{\Gamma_\alpha^s\}_{1 \leq s \leq v(G)}$, where $\Gamma_\alpha^s = \{\beta \in \Gamma : |\alpha \cap \beta| = s\}$. These sets can be bounded above by

$$|\Gamma_\alpha^s| \leq \binom{v(G)}{s} \binom{n}{v(G)-s} \leq \binom{v(G)}{s} \frac{n^{v(G)-s}}{(v(G)-s)!}.$$

Now, recalling (1.5),

$$\mathbb{E}\theta_\alpha(G) = \sum_{s=1}^{v(G)-1} \sum_{\beta \in \Gamma_\alpha^s} \sum_{G' \in R(G)} \mathbb{E}X_\alpha(G)X_\beta(G') + \sum_{\substack{G' \in R(G) \\ G \neq G'}} \mathbb{E}X_\alpha(G)X_\alpha(G').$$

To bound the expectations in the above expression, we consider the cases of different overlap s separately.

Firstly, for $G \neq G'$, and for $s = v(G)$, so that $\alpha = \beta$, there must be at least 1 edge present in G' which is not in G . Due to the conditional independence of the edges, for any edge indicator $Y_{i,j}$ which is not included in $X_\alpha(G)$,

$$\mathbb{P}(Y_{i,j} = 1 | X_\alpha(G) = 1) = \sum_{a,b=1}^Q \pi_{a,b} \mathbb{P}(i \in a, j \in b | X_\alpha(G) = 1) \leq \pi^*. \quad (2.8)$$

Hence

$$\mathbb{E}X_\alpha(G)X_\beta(G') \leq \mu(G)\pi^* \quad \text{for } \beta \in \Gamma_\alpha^v(G).$$

Next, we consider the case $s = 1$, in which α and β only intersect at a single vertex. As a result, G and G' cannot share an edge. Using the generalisation of (2.8) that for any set of edges A which does not overlap with the edges in $X_\alpha(G)$,

$$\mathbb{P}(Y_{i,j} = 1, (i,j) \in A | X_\alpha(G) = 1) \leq (\pi^*)^{|A|}, \quad (2.9)$$

it follows that

$$\mathbb{E}X_\alpha(G)X_\beta(G') \leq \mu(G)(\pi^*)^{e(G)} \quad \text{for } \beta \in \Gamma_\alpha^1.$$

Finally, we consider the case $2 \leq s \leq v(m) - 1$. We shall derive two bounds for the expectation $\mathbb{E}X_\alpha(G)X_\beta(G')$.

There are $e(G)$ edges from the subgraph G given on α and we now consider the number of additional edges resulting from the subgraph G' given on β . Here the underlying graph is K_n , the complete graph. Consider the subgraph H of the union graph of G and G' induced on the intersection of α and β . Due to the fact that $|\alpha \cap \beta| = s$, we have $v(H) = s$, and, because G is strictly balanced, it must be the case that $d(H) < d(G)$, and so $e(H) < sd(G)$. Recalling (2.2), we have $e(H) + \gamma(G) \leq sd(G)$, that is $e(H) \leq sd(G) - \gamma(G)$. Thus, there are at least $e(G) - (sd(G) - \gamma(G)) = e(G) - sd(G) + \gamma(G)$ edges from G' which are not in the subgraph G , and so the union graph of G and G' on $\alpha \cup \beta$ has at least $2e(G) - sd(G) + \gamma(G)$ edges.

Alternatively, with α as in (2.1),

$$\begin{aligned} e(G) - e(H) &= (v(G) - v(H)) \frac{e(G) - e(H)}{v(G) - v(H)} \\ &\geq (v(G) - v(H))\alpha(G) \\ &= (v(G) - s)\alpha(G), \end{aligned}$$

and therefore there are at least $e(G) + (v(G) - s)\alpha(G)$ edges in the union graph of G and G' on $\alpha \cup \beta$. This bound in connection with (2.9) leads to the bound

$$\mathbb{E}X_\alpha(G)X_\beta(G') \leq \mu(G)(\pi^*)^{\kappa(G,s)} \quad \text{for } \beta \in \Gamma_\alpha^s,$$

where $\kappa(G, s) = \max(e(G) - sd(G) + \gamma(G), (v(G) - s)\alpha(G))$. Collecting the bounds gives

$$\begin{aligned} \mathbb{E}\theta_\alpha(G) &\leq \mu(G) \left\{ \sum_{\substack{G' \in R(G) \\ G \neq G'}} \pi^* + \sum_{\substack{\beta \in \Gamma_\alpha^1 \\ G' \in R(G)}} (\pi^*)^{e(G)} + \sum_{s=2}^{v(G)-1} \sum_{\substack{\beta \in \Gamma_\alpha^s \\ G' \in R(G)}} (\pi^*)^{\kappa(G,s)} \right\} \\ &\leq \rho(G)\mu(G) \left\{ \pi^* + \frac{v(G)^2}{v(G)!} n^{v(G)-1} (\pi^*)^{e(G)} \right. \\ &\quad \left. + \sum_{s=2}^{v(G)-1} \binom{v(G)}{s} \frac{n^{v(G)-s} (\pi^*)^{\kappa(G,s)}}{(v(G)-s)!} \right\}. \end{aligned} \quad (2.10)$$

Finally, substituting (2.7) and (2.10) into (1.6) and recalling (1.4) yields (2.4). \square

Remark 2.2. 1. The stochastic block model structure enters the proof only through the expression for $\mu(G)$ as well as the bound (2.9).

2. Theorem 2.1 generalises Theorem 5.B of [5] for the Erdős-Rényi random graph model to the Stochastic block model. When we take $\pi_{a,b} = p$ for all a, b we recover the same rate of convergence as that given by Theorem 5.B of [5]. Indeed the graph combinatorics arguments in our proof are strongly related to those in the proof of Theorem 5.B of [5]. It should, however, be noted that our proof uses a local coupling approach whereas the proof in [5] uses size bias couplings.
3. To assess the behaviour of the bound it may be advantageous to use the bound $1 - e^{-\lambda} \leq \min(1, \lambda)$. Heuristically, a Poisson approximation should hold when $\mu(G)$ is small. When μ is so small that $\lambda < 1$ then the factor $1 - e^{-\lambda}$ is beneficial.
4. For a strictly balanced graph,

$$\kappa(G, s) \geq (v(G) - s)\alpha(G) > (v(G) - s)d(G) \quad (2.11)$$

for all $s = 0, \dots, v(G) - 1$. Let $\Delta\kappa = \kappa(G, s) - (v(G) - s)d(G)$. Then $\Delta\kappa > 0$. Using (2.9) we can bound $\mu(G) \leq (\pi^*)^{e(G)}$. If $n(\pi^*)^{d(G)}$ is bounded by c as $n \rightarrow \infty$ then $\lambda \leq \frac{\rho(G)}{v(G)!} c^{v(G)}$ and $n^{v(G)-s} (\pi^*)^{\kappa(G,s)} \leq c^{v(G)-s} (\pi^*)^{\Delta\kappa}$. Moreover the bound in Theorem 2.1 is then of order

$O(\min(n^{-1}, n^{-\frac{\Delta\kappa}{d(G)}}))$ as $n \rightarrow \infty$, with proportion vector f and graph G fixed.

5. Theorem 2.1 is not an asymptotic result but an explicit bound, which may or may not be small.
6. The result of Theorem 2.1 is perhaps most interesting when the limiting $Po(\lambda)$ distribution is non-degenerate in the limit $n \rightarrow \infty$. Suppose that there exist universal constants c and C such that $cn^{-1/d(G)} \leq \pi_{a,b} \leq Cn^{-1/d(G)}$ for all a, b . Then using the inequality $\frac{m^k}{k^k} \leq \binom{m}{k} \leq \frac{m^k}{k!}$, $1 \leq k \leq m$ and (1.4) we obtain

$$\frac{\rho(G)}{v(G)^{v(G)}} c^{e(G)} \leq \lambda \leq \frac{\rho(G)}{v(G)!} C^{e(G)}.$$

Moreover,

$$\begin{aligned} d_{TV}(\mathcal{L}(W), Po(\lambda)) &\leq \min \left(1, \frac{\rho(G)}{v(G)!} C^{e(G)} \right) \rho(G) \left\{ \frac{2v(G)^2}{v(G)!} C^{e(G)} n^{-1} \right. \\ &\quad \left. + Cn^{-1/d(G)} + \min(A, B) \right\}, \end{aligned} \quad (2.12)$$

where

$$\begin{aligned} A &= (1 + C^{\alpha(G)})^{v(G)-1} n^{1-\alpha(G)/d(G)}; \\ B &= C^{e(G)+\gamma(G)} (1 + C^{-d(G)})^{v(G)-1} n^{-\gamma(G)/d(G)}. \end{aligned}$$

7. If the number of hidden classes $Q = Q(n)$ grows with n , then the Poisson approximation for the distribution of W remains valid even if some of the edge probabilities $\pi_{a,b}$ are of order greater than $n^{-1/d(G)}$, provided that their respective proportions f_a and f_b are small enough.

Example 2.3. We now use (2.12) to obtain Poisson approximations for the number of copies of the following fixed graphs with $v \geq 3$ vertices in the $SBM(n, \pi, f)$ model. We consider the following strictly balanced graphs on v vertices each:

$G_{1,v}$ a tree on the v vertices, with $v - 1$ edges;

$G_{2,v}$ the cycle graph on the v vertices (with v edges);

$G_{3,v}$ the complete graph on v vertices with one edge removed;

$G_{4,v}$ K_v , the complete graph on v vertices.

In order to apply (2.12), we must compute the quantities $d(G)$, $\alpha(G)$ and $\gamma(G)$ for each graph G . These quantities are easy to compute, and the values are given in Table 1. If for a given graph G there exist universal constants c and C such that $cn^{-1/d(G)} \leq \pi_{a,b} \leq Cn^{-1/d(G)}$ for all a, b , then a bound for the total variation distance between the distribution of W and the $Po(\lambda)$ distribution now follows directly from (2.12). In Table 2, for each graph G , we give the resulting bounds on the rate of convergence in terms of n . For this rate of convergence it is assumed that the proportion vector $f = f(n)$ remains constant as $n \rightarrow \infty$, and that G does not change with n . We also give a scaling of the edge probabilities that is required to give a non-degenerate λ in the limit. This scaling is given in terms of $\pi^* = \max_{1 \leq a < b \leq Q} \pi_{a,b}$ (note that all the $\pi_{a,b}$ are of the same order). Table 2 shows that the bound on the rate of convergence for the tree graph may be considerably larger than the bound on the rate of convergence in the cycle graph.

Table 1: Values of $d(G)$, $\alpha(G)$ and $\gamma(G)$

Graph G	$d(G)$	$\alpha(G)$	$\gamma(G)$
$G_{1,v}$	$\frac{v-1}{v}$	$\frac{(v-1)-1}{v-2} = 1$	$\frac{(v-1)^2}{v} - (v-2) = \frac{1}{v}$
$G_{2,v}$	1	$\frac{v-1}{v-2}$	1
$G_{3,v}$	$\frac{(v+1)(v-2)}{2v}$	$\frac{\binom{v}{2}-1-1}{v-2} = \frac{v^2-v-4}{2(v-2)}$	1/3 if $v = 3$ and $\frac{(v+1)(v-2)}{2} - \left(\binom{v}{2} - 2\right) = 1$ if $v \geq 4$
$G_{4,v}$	$\frac{v-1}{2}$	$\frac{\binom{v}{2}-1}{v-2} = \frac{v+1}{2}$	$\frac{(v-1)v}{2} - \left(\binom{v}{2} - 1\right) = 1$

3 Subgraph counts in graph models with random edge probabilities

In this section, we consider a model in which the edge probabilities are themselves random variables. Let $I = \{u, v : 1 \leq u < v \leq n\}$ be the index set

Table 2: Scaling and bounds on the rate of convergence

Graph	Scaling	$d_{TV}(\mathcal{L}(W), Po(\lambda))$
$G_{1,v}$	$\pi^* = Cn^{-v/(v-1)}$	$O(n^{-1/(v-1)}) = O((\pi^*)^{1/v})$
$G_{2,v}$	$\pi^* = Cn^{-1}$	$O(n^{-1}) = O(\pi^*)$
$G_{3,v}$	$\pi^* = Cn^{-2v/(v+1)(v-2)}$	$O(n^{-1/2}) = O((\pi^*)^{1/3})$ if $v = 3$ and $O(n^{-2/(v-1)}) = O((\pi^*)^{(v+1)(v-2)/v(v-1)})$ if $v \geq 4$
$G_{4,v}$	$\pi^* = Cn^{-2/(v-1)}$	$O(n^{-2/(v-1)}) = O(\pi^*)$

of potential edges and for $(u, v) \in I$ let $\Theta_{u,v} = \Theta_{v,u} \in [0, 1]$ be random variables; given $\Theta_{u,v} = \theta_{u,v}$ the edge indicator $Y_{u,v}$ is Bernoulli distributed with parameter $\theta_{u,v}$. Conditional on the edge probabilities $\{\Theta_{u,v} : (u, v) \in I\}$ the edge indicator variables $\{Y_{u,v} : (u, v) \in I\}$ are assumed to be independent.

We shall assume a local dependence structure for the edge probabilities: for any $(u, v) \in I$ there is a set $B_{u,v}$ such that for any edge set \mathcal{E} , the collection of random variables $\{\Theta_{u,v} : (u, v) \in \mathcal{E}\}$ is independent of the collection of random variables $\{\Theta_{x,y} : (x, y) \in (\cup_{(u,v) \in \mathcal{E}} B_{u,v})^c\}$. Moreover, we assume that $B_{u,v}$ is of the form

$$B_{u,v} = \{(x, w) \in I : x \in M(u, v), w \in N(u, v)\}.$$

We shall often think of $N(u, v)$ as being a small set compared to $\{1, \dots, n\}$, whereas $M(u, v)$ could be a large set. We denote the least upper bound on $\{N(u, v), (u, v) \in I\}$ by g so that

$$N(u, v) \leq g$$

for all (u, v) . For independent edges, if $u < v$ we take $M(u, v) = \{u\}$ and $N(u, v) = \{v\}$ so that $B_{u,v} = \{(u, v)\}$ and $g = 1$; for graphon models, we can take $M(u, v) = \{1, \dots, n\}$ and $N(u, v) = \{u, v\}$ so that $B_{u,v} = \{(x, w) \in I : w \in \{u, v\}\}$, and $g = 2$. Other examples could include exogenous covariates such as geographic location; edge random variables could be independent if they are further than a certain geographic distance away from each other.

The dependency structure is now more involved. For $\alpha = (\alpha_1, \dots, \alpha_{v(G)})$ let $\mathcal{E}(\alpha) = \{(i, j) : i \neq j, i, j \in \{\alpha_1, \dots, \alpha_{v(G)}\}\}$ denote the set of edges of the complete graph on α . Then the set

$$A_\alpha = \{\beta \in \Gamma : |\mathcal{E}(\beta) \cap (\cup_{(u,v) \in \mathcal{E}(\alpha)} B_{u,v})| \geq 1\}. \quad (3.1)$$

is a dependency neighbourhood of α . In particular, if $\beta \notin A_\alpha$ then $\{\Theta_{x,y} : (x,y) \in \mathcal{E}(\beta)\}$ is independent of $\{\Theta_{u,v} : (u,v) \in \mathcal{E}(\alpha)\}$. We can bound the size of this dependency neighbourhood as follows. For $\beta \in A_\alpha$ at least one of the vertices of β is in a set $N(u,v)$ for some $u,v \in \mathcal{E}(\alpha)$. Each of these sets $N(u,v)$ has at most g elements. Hence

$$|A_\alpha| \leq gv(G) \binom{n}{v(G)-1}. \quad (3.2)$$

For a set of edges $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_k\}$ we introduce the notation $V(\alpha)$ for the set of vertices which are endpoints in α , so that $|V(\alpha)| \leq 2|\alpha|$. We let

$$\nu_{k,v,s} = \max_{\substack{\alpha=\{\alpha_1,\alpha_2,\dots,\alpha_k\}; \beta=\{\beta_1,\beta_2,\dots,\beta_v\}: \\ \alpha \cap \beta = \emptyset; |V(\alpha) \cap V(\beta)|=s}} \mathbb{P} \left(\prod_{i=1}^k Y_{\alpha_i} = 1 \mid \prod_{j=1}^v Y_{\beta_j} = 1 \right). \quad (3.3)$$

With $\mu(G) = \mathbb{E}X_\alpha(G)$ and

$$\lambda := \mathbb{E}W = \binom{n}{v(G)} \rho(G) \mu(G) \quad (3.4)$$

we obtain the following generalisation of Theorem 2.1.

Theorem 3.1. *Assume that the $\pi_{a,b}$ are arbitrary random variables supported on a subset of $[0, 1]$. Let $\nu_{k,v,s}$ be as in (3.3). Suppose that G is a strictly balanced graph. Then*

$$\begin{aligned} d_{TV}(\mathcal{L}(W), Po(\lambda)) &\leq (1 - e^{-\lambda}) \rho(G) g \left\{ 2 \frac{v(G)^2}{v(G)!} n^{v(G)-1} \nu_{e(G), e(G), 1} + \nu_{1, e(G), 1} \right. \\ &\quad \left. + \sum_{s=2}^{v(G)-1} \binom{v(G)}{s} \frac{n^{v(G)-s} \nu_{\kappa(G,s), e(G), s}}{(v(G)-s)!} \right\}, \end{aligned} \quad (3.5)$$

where $\kappa(G, s)$ is as in Theorem 2.1.

Proof. The proof proceeds almost exactly as that of Theorem 2.1. The combinatorial arguments are exactly as before, although note the additional factor of g in (3.2). We also deal with the expectations in the formulas for $\mathbb{E}\eta_\alpha(G)$ similarly. A complication arises from bounding the expressions $\mathbb{E}X_\alpha(G)X_\beta(G')$ which occur in $\mathbb{E}\theta_\alpha(G)$; the analog of (2.9) is that for any set of edges A such that $|v(A) \cap v(G)| = s$,

$$\mathbb{P}(Y_{i,j} = 1, (i,j) \in A \mid X_\alpha(G) = 1) \leq \nu_{|A|, e(G), s}.$$

□

Remark 3.2. *In the case that the edges are independent and $\nu = \max_{\alpha} \mathbb{E}(Y_{\alpha})$, we find that $\nu_{k,v,s} = \nu^k$ does not depend on v or s . It is now an immediate consequence of Theorem 3.1 that*

$$d_{TV}(\mathcal{L}(W), Po(\lambda)) \leq (1 - e^{-\lambda})\rho(G) \left\{ 2 \frac{v(G)^2}{v(G)!} n^{v(G)-1} \nu^{e(G)} + \nu \right. \\ \left. + \sum_{s=2}^{v(G)-1} \binom{v(G)}{s} \frac{n^{v(G)-s} \nu^{\kappa(G,s)}}{(v(G)-s)!} \right\}. \quad (3.6)$$

Taking the $\pi_{a,b}$ to be constants in (3.6) yields $\pi^* = \nu$ and recovers the bound (2.4).

4 Subgraph counts in a graphon model

The h -graphon model uses

$$\pi_{u,v} = h(U_u, U_v)$$

where $h : [0, 1]^2 \rightarrow [0, 1]$ is a symmetric, measurable function and U_a , $a = 1, \dots, n$, are independent $U[0, 1]$ variables which index the graphon; see for example [1, 6, 15, 21], and [15, 26] for graphon estimation. In this case edges are not independent, but edges which do not share a vertex are independent, and we can choose $M(u, v) = \{1, \dots, n\}$ and $N(u, v) = \{u, v\}$ so that $g = 2$. Hence

$$\mu(G) = \int_{[0,1]^{v(G)}} du_1 \cdots du_{v(G)} \prod_{1 \leq i < j \leq v(G): (i,j) \in E(G)} h(u_i, u_j). \quad (4.1)$$

The weak dependence structure yields the following theorem.

Theorem 4.1. *Let $\pi_{u,v} = h(U_u, U_v)$ where $h : [0, 1]^2 \rightarrow [0, 1]$ is a symmetric, measurable function and U_a , $a = 1, \dots, n$, are independent $U[0, 1]$ variables and let*

$$h^* = \max_{u,v} h(u, v).$$

Suppose G is a strictly balanced graph. Then

$$d_{TV}(\mathcal{L}(W), Po(\lambda)) \leq 2(1 - e^{-\lambda})\rho(G) \left\{ 2 \frac{v(G)^2}{v(G)!} n^{v(G)-1} (h^*)^{e(G)} + h^* \right. \\ \left. + \sum_{s=2}^{v(G)-1} \binom{v(G)}{s} \frac{n^{v(G)-s} (h^*)^{\kappa(G,s)}}{(v(G)-s)!} \right\},$$

where $\kappa(G, s)$ is given in (2.5).

Proof. We partition A_α , as defined in (3.1), into sets $\{\Gamma_\alpha^s\}_{1 \leq s \leq v(G)}$, where $\Gamma_\alpha^s = \{\beta \in \Gamma : |\alpha \cap \beta| = s\} \cap A_\alpha$. Recalling (1.5), we have

$$\mathbb{E}\theta_\alpha(G) = \sum_{s=1}^{v(G)-1} \sum_{\beta \in \Gamma_\alpha^s} \sum_{G' \in R(G)} \mathbb{E}X_\alpha(G)X_\beta(G') + \sum_{\substack{G' \in R(G) \\ G \neq G'}} \mathbb{E}X_\alpha(G)X_\alpha(G').$$

Firstly, for $G \neq G'$, and for $s = v(G)$, so that $\alpha = \beta$, there must be at least 1 edge present in G' which is not in G . Due to the conditional independence of the edges, for any edge indicator $Y_{i,j}$ which is not included in $X_\alpha(G)$,

$$\begin{aligned} \mathbb{P}(Y_{i,j} = 1 | X_\alpha(G) = 1) &= \int_{[0,1]^{v(G)}} du_1 \cdots du_{v(G)} \mathbb{P}(Y_{i,j} = 1 | U_v = u_v, v \in V(G)) \\ &= \int_{[0,1]^{v(G)}} du_1 \cdots du_{v(G)} h(u_i, u_j) \\ &\leq h^*. \end{aligned} \tag{4.2}$$

Hence

$$\mathbb{E}X_\alpha(G)X_\beta(G') \leq \mu(G)h^* \quad \text{for } \beta \in \Gamma_\alpha^v(G).$$

Next, we consider the case $s = 1$, in which α and β only intersect at a single vertex. As a result, G and G' cannot share an edge. Using the generalisation of (2.8) that for any set of edges A which does not overlap with the edges involved in $X_\alpha(G)$,

$$\mathbb{P}(Y_{i,j} = 1, (i, j) \in A | X_\alpha(G) = 1) \leq (h^*)^{|A|}, \tag{4.3}$$

we obtain that

$$\mathbb{E}X_\alpha(G)X_\beta(G') \leq \mu(G)(h^*)^{e(G)} \quad \text{for } \beta \in \Gamma_\alpha^1.$$

Finally, we consider the case $2 \leq s \leq v(m) - 1$. As in the proof of Theorem 2.1, the union graph of G and G' on $\alpha \cup \beta$ has at least $\kappa(G, s)$ edges. This bound in connection with (4.3) leads to the bound

$$\mathbb{E}X_\alpha(G)X_\beta(G') \leq \mu(G)(h^*)^{\kappa(G,s)} \quad \text{for } \beta \in \Gamma_\alpha^s.$$

Collecting the bounds gives the assertion as in Theorem 2.1. \square

Remark 4.2. *In the proof of Theorem 4.1 we could have replaced (4.2) by*

$$\begin{aligned} \mathbb{P}(Y_{i,j} = 1 | X_\alpha(G) = 1) &= \int_{[0,1]^{v(G)}} du_1 \cdots du_{v(G)} h(u_i, u_j) \\ &\leq \mathbb{E} \left[\max_{U_i, U_j: i \neq j \in v(G)} h(U_i, U_j) \right]. \end{aligned} \quad (4.4)$$

For example, if $h(x, y) = \frac{1}{2}(x + y)$ then $h^* = 1$ whereas, using the order statistic notation,

$$\mathbb{E} \left[\max_{U_i, U_j: i \neq j \in v(G)} h(U_i, U_j) \right] = \frac{1}{2} \mathbb{E}(U_{(n)} + U_{(n-1)}) = \frac{2v(G) - 1}{2(v(G) + 1)} < 1.$$

Similarly, (4.3) could be replaced by

$$\mathbb{P}(Y_{i,j} = 1, (i, j) \in A | X_\alpha(G) = 1) \leq \mathbb{E} \left[\max_{U_i, i \in v(G)} \prod_{(i,j) \in A} h(U_i, U_j) \right]. \quad (4.5)$$

While (4.4) and (4.5) would yield numerically smaller bounds, the order of the bounds would not be affected unless $v(G)$ depends on n . In contrast, h^* is easier to calculate in applications.

Example 4.3. *In analogy to copulas, where Archimedean copulas have proved a useful concept, consider what can be coined an Archimedean graphon: Let $h : [0, 1]^2 \rightarrow [0, 1]$ be given by $h(x, y) = \psi(\psi^{[-1]}(x) + \psi^{[-1]}(y))$ where $\psi : [0, \infty) \rightarrow [0, 1]$ is a continuous, strictly decreasing function which is convex on the open interval $(0, \infty)$ and $\psi^{[-1]}(x) = \inf\{u : \psi(u) \leq x\}$ is its generalised inverse. Using the Williamson transform we can write*

$$\psi(x) = \int_{(x, \infty)} \left(1 - \frac{x}{t}\right) dF_R(t) = \mathbb{E} \left(1 - \frac{x}{R}\right)_+,$$

where F_R is the c.d.f. of a non-negative random variable R which has no atom at zero, see for example [18]. If $\inf\{x : dF_R(x) > 0\} = a_R$ with $a_R > 0$ then

$$h^* \leq \sup_{x \geq 0} \psi(x) = \int_{a_R}^{\infty} \left(1 - \frac{a_R}{t}\right) dF_R(t) = 1 - a_R \mathbb{E}(R^{-1}).$$

In contrast, $\mathbb{E} \left[\min_{U_i, i \in v(G)} \prod_{(i,j) \in A} \psi(\psi^{[-1]}(U_i) + \psi^{[-1]}(U_j)) \right]$ as used in (4.5) would be more difficult to calculate.

The next example illustrates how scaling considerations enter in the distributional bound.

Example 4.4. Let $h : [0, 1]^2 \rightarrow [0, 1]$ be given by $h(x, y) = xy$. In this case, (4.1) gives that

$$\mu(G) = \int_{[0,1]^{v(G)}} du_1 \cdots du_{v(G)} \prod_{i \in V(G)} u_i^{\deg_G(i)} = \prod_{i \in V(G)} \frac{1}{\deg_G(i) + 1},$$

where $\deg_G(i)$ is the degree of i in G , that is, the number of edges in $E(G)$ which have i as an end point; $1 \leq \deg_G(i) \leq v(G) - 1$. Thus in order to obtain a moderate value of λ , the graph G has to have a large number of vertices with degrees which typically grow like n ; such graphs are also called dense graphs. In this example, $h^* = 1$ and the bound in Theorem 4.1 will be of the order $n^{v(G)}$ if the graph G is fixed.

If instead we consider the function $f_n : [0, 1]^2 \rightarrow [0, 1]$; $h_n(x, y) = n^{-\frac{1}{d(G)}} xy$ then the limiting Poisson distribution is not-degenerate and as in (2.12) the bound in Theorem 4.1 tends to 0 with n tending to ∞ .

Finally, we note that the h -graphon model can be viewed as a stochastic block model if h is piecewise constant. If $0 = s_1 < s_2 < \cdots < s_{Q-1} = 1$, where $s_i = \sum_{k=1}^i f_k$, is a partition of $[0, 1]$ so that h is constant on each rectangle $[s_i, s_{i+1}) \times [s_j, s_{j+1})$, then we could assign type i to vertex v if $U_v \in [s_i, s_{i+1})$. The randomness now lies only in the class assignments. In this case we recover Theorem 2.1.

Acknowledgements

MC acknowledges support from the Department of Statistics, University of Oxford, for a Summer Studentship. RG and GR acknowledge support from EPSRC grant EP/K032402/1.

References

- [1] Airoldi, E. M., Costa, T. B. and Chan, S. H. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. *Adv. Neur. In.* **26** (2013), pp. 692–700.
- [2] Aldous, D. J. Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** (1981), pp. 581–598.
- [3] Ali W., Rito, T., Reinert, G., Sun, F. and Deane, C. M. Alignment-free protein interaction network comparison. *Bioinformatics* **30** (2014), pp. i430–i437.
- [4] Arratia, R. Goldstein, L. and Gordon, L. Two Moments Suffice for Poisson Approximations: the Chen-Stein Method. *Ann. Probab.* **17** (1989), pp. 9–25.
- [5] Barbour, A. D., Holst, L. and Janson, S. *Poisson Approximation*. Oxford University Press, Oxford, 1992.
- [6] Bickel, P. and Chen, A. A non parametric view of network models and Newman-Girvan and other modularities. *P. Natl. Acad. Sci. USA* **106** (2009), pp. 21068–21073.
- [7] Bollobas, B., Janson, S. and Riordan, O. The phase transition in inhomogeneous random graphs. *Random Struct. Algor.* (2007), **31**, pp. 3–122.
- [8] Chen, L. H. Y. Poisson approximation for dependent trials. *Ann. Probab.* **3** (1975), pp. 534–545.
- [9] Condon, A. and Karp, R. M. Algorithms for graph partitioning on the planted partition model. In *Randomization, Approximation, and Combinatorial Optimization. Algorithms and Techniques* Springer Berlin Heidelberg, (1999), pp. 221–232.
- [10] Daudin, J. J., Picard, F. and Robin, S. A mixture model for random graphs. *Stat. Comput.* **18** (2008), pp. 173–183.
- [11] Diaconis, P. and Janson, S. Graph limits and exchangeable random graphs. *Rendiconti di Matematica* **28** (2008), pp. 33–61.

- [12] Frank, O. and Strauss, D. Markov graphs. *J. Am. Stat. Assoc.* **81** (1986), pp. 832–842.
- [13] Holland, P. W., Laskey, K. B. and Leinhardt, S. Stochastic blockmodels: First steps. *Soc. networks* **5** (1983), pp. 109–137.
- [14] Karrer, B., and Newman, M. E. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**(1) (2011), 016107.
- [15] Latouche, P. and Robin, S. Bayesian Model Averaging of Stochastic Block Models to Estimate the Graphon Function and Motif Frequencies in a W-graph Model. arXiv:1310.6150, 2013.
- [16] Lovász, L. and Szegedy, B. Limits of dense graph sequences. *J. Comb. Theory B* **96** (2006), pp. 933–957.
- [17] Matias, C. and Robin, S. Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proceedings and Surveys* **47** (2014), pp. 55–74.
- [18] McNeil, A. and Neslehová, J. Multivariate Archimedean Copulas, d-monotone functions and L_1 -norm symmetric distributions. *Ann. Stat.* **37** (2007), pp. 3059–3097.
- [19] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. Network motifs: simple building blocks of complex networks. *Science* **298** (2002), pp. 824–827.
- [20] Nowicki, K. and Snijders, T. Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.* **96** (2001), pp. 1077–1087.
- [21] Olhede, S. C. and Wolfe, P. J. Network histograms and universality of blockmodel approximation. *P. Natl. Acad. Sci. USA* **111** (2014), pp. 14722–14727.
- [22] Picard, F., Daudin, J. J., Koskas, M., Schbath, S. and Robin, S. Assessing the exceptionality of network motifs. *J. Comput. Biol.* **15** (2008), pp. 1–20.
- [23] Ruciński, R. J. and Vince, A. Balanced graphs and the problem of subgraphs of random graphs. *Congressus Numerantium* **49** (1985), pp. 181–190.

- [24] Sarajlić, A., Janjić, V., Stojković, N., Radak, D. and Pržulj, N. Network topology reveals key cardiovascular disease genes. *PLoS ONE* **8**(8) (2013), e71537.
- [25] Stark, D. Compound Poisson approximation of subgraph counts in random graphs. *Random Struct. Algor.* **18** (2001), pp. 39–60.
- [26] Wolfe, P. J. and Olhede, S. C. Nonparametric graphon estimation. arXiv:1309.5936, 2013.