Adv. Appl. Prob. **50,** 759–782 (2018)
*doi:*10.1017/apr.2018.35
© *Applied Probability Trust* 2018

COMPOUND POISSON APPROXIMATION OF SUBGRAPH COUNTS IN STOCHASTIC BLOCK MODELS WITH MULTIPLE EDGES

MATTHEW COULSON,* *University of Birmingham*
ROBERT E. GAUNT,** *The University of Manchester*
GESINE REINERT,*** *University of Oxford*

Abstract

We use the Stein–Chen method to obtain compound Poisson approximations for the distribution of the number of subgraphs in a generalised stochastic block model which are isomorphic to some fixed graph. This model generalises the classical stochastic block model to allow for the possibility of multiple edges between vertices. We treat the case that the fixed graph is a simple graph and that it has multiple edges. The former results apply when the fixed graph is a member of the class of strictly balanced graphs and the latter results apply to a suitable generalisation of this class to graphs with multiple edges. We also consider a further generalisation of the model to pseudo-graphs, which may include self-loops as well as multiple edges, and establish a parameter regime in the multiple edge stochastic block model in which Poisson approximations are valid. The results are applied to obtain Poisson and compound Poisson approximations (in different regimes) for subgraph counts in the Poisson stochastic block model and degree corrected stochastic block model of Karrer and Newman (2011).

Keywords: Stochastic block model; multiple edges; subgraph count; compound Poisson approximation; Stein–Chen method; pseudo-graph

2010 Mathematics Subject Classification: Primary 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 [16]. Furthermore, many networks are conjectured to have characteristic over- or under-represented motifs (small subgraphs); see, for example, [22]. Moreover, statistics based on small subgraph counts can be used to compare networks, as shown in [29] and [32]. To determine which small subgraphs are unusual, assessing the distribution of such motifs is key. The mean and variance of subgraph counts for some common random graph models can be found in [26], whilst the focus of this paper concerns distributional approximation of subgraph counts.

A powerful and popular tool for deriving distributional approximations for subgraph counts is the Stein–Chen method; see [11] and [31]. The method was first used to derive approximations for subgraph counts in [3] and a comprehensive account of Poisson approximation of subgraph counts can be found in [10]. To date, distributional approximations for subgraph counts

Received 19 October 2017; revision received 20 June 2018.
* Postal address: School of Mathematics, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK.
** Postal address: School of Mathematics, The University of Manchester, Manchester M13 9PL, UK.
Email address: robert.gaunt@manchester.ac.uk
*** Postal address: Department of Statistics, University of Oxford, 24–29 St Giles', Oxford OX1 3LB, UK.

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

have mostly been derived for simple random graph models (simple graphs are graphs with no self-loops or multiple edges). However, networks with multiple edges (often referred to as multigraphs) arise naturally in many real-world networks and also in important synthetic networks models, such as the configuration model. For example, in collaboration networks as analysed in [17], vertices are authors, and two authors are linked by an edge if they have co-authored at least one paper. One can argue that edges in collaboration networks simplify relationships and that the number of joint papers contains important information. Another example is the Molloy–Reed construction of realisations from the configuration model [23] which allows self-loops as well as multiple edges. A collection of currently 90 networks with multi-edges can be found in the KONECT data base; available at http://konect.uni-koblenz.de/. In this paper we derive distributional approximations for subgraph counts in a generalised stochastic block model, which includes the Poisson stochastic block model and the degree-corrected stochastic block model of [19] as special cases.

The stochastic block model (SBM) was introduced originally for directed graphs by Holland *et al.* [18] and generalised to other graphs by Nowicki and Snijders [24]. It is also called the Erdős–Rényi mixture model in [15] and in theoretical computer science it is known as the planted partition model; see [12]. The SBM has a wide range of applications; see the survey [21] and the references therein. In this paper we will use the following generalisation of the classical SBM to one with multiple edges. Consider an undirected random graph on n vertices, with no self-loops, in which the vertices are spread among Q hidden classes. Letting $f = (f_1, \dots, f_Q)$ be a vector in $[0, 1]^Q$ such that $\sum_{q=1}^{Q} f_q = 1$, the class label of a vertex is drawn from a multinomial distribution $\mathcal{M}(1, f)$, and class assignments are independent of each other. The number of edges $\tilde{Y}_{i,j}$ between vertices i and j (the tilde in the notation distinguishes these random variables from the random variables $Y_{a,b}$ defined in (2.6)) are independent conditionally on the class of the vertices, and the probability depends only on the classes of the vertices: denoting by $c_v \in \{1, \dots, Q\}$ the class of vertex i,

$$\mathbb{P}(\tilde{Y}_{i,j} = k \mid c_i = a, c_j = b) = \pi_{a,b,k}, \qquad k = 0, 1, 2, \dots, \tag{1.1}$$

and $\sum_{k=0}^{\infty} \pi_{a,b,k} = 1$ for all a and b. We will refer to this model as a *stochastic block multigraph model* (SBMM) and denote it by SBMM(n, π, f). The model of [19] corresponds to the case of Poisson probabilities $\pi_{a,b,k} = \mathrm{e}^{-\omega_{a,b}} \omega_{a,b}^k / k!$. If $\pi_{a,b,k} = 0$ for $k \geq 2$ then the SBMM model reduces to the classical SBM. Further, if

$$\pi_{a,b,0} = 1 - p, \qquad \pi_{a,b,1} = p, \qquad \pi_{a,b,k} = 0 \quad \text{for all } k \geq 2,$$

the SBM reduces to the Erdős–Rényi random graph model, which we denote by $\mathfrak{G}(n, p)$. In Section 5 we will also introduce a generalisation of the SBMM to pseudo-graphs (graphs which allow for self-loops as well as multiple edges). In this paper we assume that π and f are known, and that $f_1, \dots, f_Q > 0$; for estimating these quantities for the classical SBM, see, for example, [1], [20], [21], and [25].

In most of this paper we will focus on the case that vertices in the same class are stochastically equivalent, that is, the probabilities are given by (1.1). However, as noted by Karrer and Newman [19], such models may be a poor fit when modelling networks with hubs or highly varying degrees within communities. Indeed, Karrer and Newman [19] introduced the degree-corrected SBM in which the probabilities are given by $\pi_{a,b,i,j,k} = \mathrm{e}^{-\theta_i \theta_j \omega_{a,b}} (\theta_i \theta_j \omega_{a,b})^k / k!$, where θ_i relates to an expected degree of vertex i. All our results and proofs generalise easily

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

to such heterogeneous vertices, in which the right-hand side of (1.1) is of the form $\pi_{a,b,i,j,k}$; see Corollary 3.1 and Example 3.3.

In [13], it was shown that the distribution of the number copies of a fixed graph G in the SBM(n, π, f) model is well approximated by an appropriate Poisson distribution if G is a member of the class of strictly balanced graphs (defined in Section 2.1.1) as long as $\pi_{a,b,1}$ is not too large for all a and b. Explicit bounds for the rate of convergence (in the total variation distance) can also be found in [13]. The results of [13] generalise well-known results for the $\mathfrak{G}(n, p)$ model, such as [10, Theorem 5.B]. For normal approximations of subgraph counts there are results available for the $\mathfrak{G}(n, p)$ model as well as for the configuration model; see [6] and the references therein.

In this paper we generalise the results of [13] to derive compound Poisson approximations for subgraph counts in the SBMM(n, π, f) model. We consider both the case that the fixed graph G is a simple graph and that it is itself a multigraph. For simple graphs, a Poisson approximation is valid when the fixed graph G is strictly balanced; see, for example, [10, Theorem 5.B]. For the case of multigraphs we introduce a related notion which we call 'strictly pseudo-balanced'; see Definition 2.1. The existence of multiple edges can create clumps of subgraph counts. In such a situation, a compound Poisson approximation is more appropriate than a Poisson approximation; see also Remark 2.1 for why this is the case. The compound Poisson approximations of the main results (Theorems 3.1 and 4.1) are valid when the fixed graph G is strictly balanced (or strictly pseudo-balanced) and the edge probabilities $\pi_{a,b,k}$ are not too large for all a, b and k, and thus generalise [13, Theorem 2.1] in a natural manner. In certain parameter regimes, which include the model of [19], not only a compound Poisson approximation but also a Poisson approximation may hold; see Section 5. To the best of the authors' knowledge, this is the first paper to contain explicit bounds for a compound Poisson approximation of subgraph counts in random multigraph models.

The rest of the paper is organised as follows. In Section 2 we introduce the setup we use to prove our main results. Section 2.1 contains the notation and procedure we use to count subgraph copies in the SBMM, while in Section 2.2 we recall the relevant theory from the Stein–Chen method for compound Poisson approximation that will be used to derive the approximations. Section 2.3 concerns the calculation of the parameter in the limiting compound Poisson distribution. As an example, the parameter is calculated in the case that the fixed graph G is the complete graph on v vertices. In Section 3 we use the Stein–Chen method to derive a compound Poisson approximation for the number of subgraphs in the SBMM which are isomorphic to some fixed simple graph from the class of strictly balanced graphs. In Remark 2.1 we illustrate with an example that a Poisson approximation is, in general, not applicable. We note in Example 3.2 that Theorem 3.1 is easily applicable to the model of [19]. Also, in Corollary 3.1 and Example 3.3, we show how Theorem 3.1 generalises to edge probabilities of the form $\pi_{a,b,i,j,k}$, which includes the degree-corrected SBM of [19] as a special case. Section 4 is devoted to a generalisation of this problem in which the fixed graph can now itself have multiple edges. We again derive a compound Poisson approximation, for which the strictly balanced graph condition is replaced by an appropriate generalisation for multigraphs. Section 5 contains the generalisation of our results to the SBMM with self-loops. We obtain some simple sufficient conditions for which a Poisson approximation of subgraph counts holds in the SBMM; see Theorem 5.2 and Remark 5.2. Finally, in Corollary 5.1, we present a Poisson approximation for subgraph counts in the Poisson SBM of [19].

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

2. The setup

2.1. Subgraph counts in the SBMM

2.1.1. *Simple graph G.* We introduce some notation that will be useful for counting copies of a fixed graph G in the SBMM(n, π, f). For simplicity, we first consider the case that G is a simple graph and then in Section 2.1.2 we generalise to the case that G is a multigraph. Let K_n be the complete graph with n vertices and $\binom{n}{2}$ edges. Let $G = (V(G), E(G)) \subset K_n$ be a fixed graph with $v(G)$ vertices and $e(G)$ edges; here $V(G)$ denotes its vertex set and $E(G)$ its edge set. To avoid trivialities, we assume that $e(G) \geq 1$ and that G has no isolated vertices. We are particularly interested in the case that G is a member of the class of strictly balanced graphs, which is defined as follows; see, for example, [10]. Let

$$d(G) = \frac{e(G)}{v(G)}. \tag{2.1}$$

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

Let Γ be the set of all copies of G in K_n, that is, all subgraphs of K_n that are isomorphic to G. Note that $|\Gamma| = \binom{n}{v(G)}\rho(G)$, where $\binom{n}{v(G)}$ is the number of ways of choosing $v(G)$ vertices, and

$$\rho(G) = \frac{(v(G))!}{a(G)}, \tag{2.2}$$

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

In a multigraph, more than one isomorphic copy of G may be present for a given labelled graph $\alpha = (V(\alpha), E(\alpha)) \in \Gamma$. To take this into account, for $\alpha \in \Gamma$ with $E(\alpha) = \{(\alpha_1, \alpha_2, \ldots, \alpha_{e(G)})\}$, let $\Lambda_\alpha = \Lambda_\alpha(G)$ denote the set $\{(a_{\alpha_1}, \ldots, a_{\alpha_{e(G)}}) \in \{1, 2, 3, \ldots\}^{e(G)}\} = \mathbb{N}^{e(G)}$. We view the a_{α_i} as edge labels in the following sense. In the SBMM(n, π, f) model, the number of edges between vertices u and v is determined according to the classes of these vertices and the measure π. If $k \geq 1$ edges are selected via the measure π, then these edges are given edge labels $1, \ldots, k$. In this way, if vertices u and v are from classes a and b, then edge number 1 between u and v occurs with probability $\sum_{k=1}^{\infty} \pi_{a,b,k}$, edge number 2 between u and v occurs with probability $\sum_{k=2}^{\infty} \pi_{a,b,k}$, and so forth.

Now, let $\mathfrak{G} = (\mathcal{V}, \mathcal{E})$ be a random multigraph on n edges. For $\alpha \in \Gamma$ and $a \in \Lambda_\alpha$, let $X_{\alpha,a}(G)$ be the indicator random variable for the occurrence of $\alpha \in \Gamma$ which is isomorphic to the fixed graph G, using edges $a \in \Lambda_\alpha$. Here we use the shorthand 'using edges $a \in \Lambda_\alpha$' to indicate that the occurrence is on edge labelled a_{α_1} of edge α_1, on edge labelled a_{α_2} of edge α_2, and so on, until edge labelled $a_{\alpha_{e(G)}}$ of edge $\alpha_{e(G)}$. Let $W(G) = W$ denote the total number of copies of G in the random graph $\mathfrak{G}$,

$$W(G) = W = \sum_{\alpha \in \Gamma} \sum_{a \in \Lambda_\alpha} X_{\alpha,a}(G). \tag{2.3}$$

In the following, the dependence of W and $X_{\alpha,a}(G)$ on G is usually suppressed to simplify notation. 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 $\frac{1}{2}(n-1)!$ copies, but no induced copy, of an n-cycle.

As an illustration of this notation, count the number of isomorphic copies of the path on three vertices, denoted by G, in a graph $\mathfrak{G}$ with vertex set $\{1, 2, 3\}$. Suppose that $\mathfrak{G}$ has a single edge

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

between vertices 1 and 2, three edges between vertices 2 and 3, and no edge between vertices 1 and 3. Then the set of possible copies of G in K_3 is

$$\begin{aligned}\Gamma &= \{(\{1,2,3\},\{(1,2),(2,3)\}),(\{1,2,3\},\{(1,3),(2,3)\}),(\{1,2,3\},\{(1,2),(1,3)\})\}\\ &= \{\alpha^{(1)},\alpha^{(2)},\alpha^{(3)}\}.\end{aligned}$$

In this example, $X_{\alpha^{(1)},(1,1)} = X_{\alpha^{(1)},(1,2)} = X_{\alpha^{(1)},(1,3)} = 1$ and all other indicators are equal to 0. Thus, there are three copies of G in $\mathfrak{G}$ and, hence, $W(G) = 3$.

In the SBMM(n,π,f), the conditional occurrence probability of an isomorphic copy of the subgraph G on

$$\alpha = (\{i_1,\dots,i_{v(G)}\},\{\alpha_1,\dots,\alpha_{e(G)}\}) \in \Gamma$$

with edge multiplicities $(a_{\alpha_1},\dots,a_{\alpha_{e(G)}}) \in \mathbb{N}^{e(G)}$, given the classes $c_{i_1},\dots,c_{i_{v(G)}}$ of the vertices $i_1,\dots,i_{v(G)}$ forming α, is

$$\mathbb{P}(X_{\alpha,a} = 1 \mid c_{i_1},\dots,c_{i_{v(G)}}) = \prod_{1\le u<v\le v(G):\,(u,v)\in E(\alpha)} q_{c_u,c_v,a_{(u,v)}},$$

where $q_{a,b,k} = \sum_{l=k}^{\infty}\pi_{a,b,l}$ is the probability of there being at least k edges between vertices of classes a and b. Here $a_{(u,v)}$ denotes the edge label $a_{(u,v)}$ between vertices u and v. The occurrence probability of an isomorphic copy of G, using edges $a \in \Lambda_\alpha$, is

$$\mu_{\alpha,a}(G) = \mathbb{E}X_{\alpha,a} = \sum_{c_1,c_2,\dots,c_{v(G)}=1}^{Q} f_{c_1}f_{c_2}\cdots f_{c_{v(G)}} \prod_{1\le u<v\le v(G):\,(u,v)\in E(\alpha)} q_{c_u,c_v,a_{(u,v)}},$$

and the expected number of isomorphic copies of G at position $\alpha \in \Gamma$ is, therefore,

$$\begin{aligned}\mu(G) &= \sum_{a\in\Lambda_\alpha}\mathbb{E}X_{\alpha,a} \qquad (2.4)\\ &= \sum_{a\in\Lambda_\alpha}\sum_{c_1,c_2,\dots,c_{v(G)}=1}^{Q} f_{c_1}f_{c_2}\cdots f_{c_{v(G)}} \prod_{1\le u<v\le v(G):\,(u,v)\in E(\alpha)} q_{c_u,c_v,a_{(u,v)}}\\ &= \sum_{c_1,c_2,\dots,c_{v(G)}=1}^{Q} f_{c_1}f_{c_2}\cdots f_{c_{v(G)}} \prod_{1\le u<v\le v(G):\,(u,v)\in E(\alpha)} \Bigg(\sum_{a_{(u,v)}=1}^{\infty} q_{c_u,c_v,a_{(u,v)}}\Bigg)\\ &= \sum_{c_1,c_2,\dots,c_{v(G)}=1}^{Q} f_{c_1}f_{c_2}\cdots f_{c_{v(G)}} \prod_{1\le u<v\le v(G):\,(u,v)\in E(\alpha)} \mathbb{E}Y_{c_u,c_v}, \qquad (2.5)\end{aligned}$$

where the random variable $Y_{a,b}$ has probability mass function

$$\mathbb{P}(Y_{a,b} = k) = \pi_{a,b,k}, \qquad k = 0,1,2,\dots. \qquad (2.6)$$

We write $\mu(G)$ (with no α subscript) in (2.4), since $\mu(G)$ is constant over $\alpha \in \Gamma$, for a fixed Γ, since all the graphs in Γ are copies of the same graph. From (2.3) and (2.6), it follows that

$$\mathbb{E}W = \binom{n}{v(G)}\rho(G)\mu(G). \qquad (2.7)$$

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

2.1.2. *Multigraph G*. We now generalise the setup for simple fixed graphs G to fixed multigraphs G. Assume that the graph G has $v(G)$ vertices and denote the maximum number of edges between any two nodes by the finite number $t(G)$. As before, the vertex and edge sets are denoted by $V(G)$ and $E(G)$, respectively. Here $E(G)$ differs from the simple graph case since there may now be more than one edge between a pair of vertices. Multiple edges appear in $E(G)$ according to their multiplicities in G. Again, assume that there are no isolated vertices and the total number of edges $e(G)$ is greater or equal to 1. The notion of strictly balanced graphs has the following natural generalisation to multigraphs.

Definition 2.1. Let $f(G)$ be the number of pairs of vertices in G with at least one edge between them and set

$$\partial(G) = \frac{f(G)}{v(G)}. \tag{2.8}$$

Then we say that G is a member of the class of *strictly pseudo-balanced multigraphs* if $\partial(H) < \partial(G)$ for all subgraphs $H \subsetneq G$.

Note that condition (2.8) is equivalent to requiring the subgraph obtained by reducing the multiplicity of all edges to 1 to be strictly balanced.

The procedure for counting copies of G is similar to the simple graph case, but with some important generalisations. The quantity $\rho(G)$ is defined as in (2.2). It is worth noting, for example, that if G_1 is a triangle then $\rho(G_1) = 1$, whereas for the multigraph G_2 which is a triangle with an additional edge added between a pair of vertices then $\rho(G_2) = 3$. Note that there is a discrepancy here even though the cardinality of the automorphism groups would be equal if the multiplicity of the multi-edge in G_2 was set to 1. Now, Γ is the set of all isomorphic copies of G in the complete $t(G)$-multigraph on n vertices, for which there are $t(G)$ edges between all $\binom{n}{2}$ vertex pairs. As before, $|\Gamma| = \rho(G)\binom{n}{v(G)}$. Note that for counting copies of G_1, we have $|\Gamma| = \binom{n}{3}$, but for counting copies of G_2, the set Γ has a larger cardinality: $|\Gamma| = 3\binom{n}{3}$.

We are also required to suitably generalise the set Λ_α. To explain the generalisation, first suppose that G consists of only one edge with multiplicity k. Then $\alpha \in \Gamma$ consists of k edges between a particular pair of vertices. If the observed count is $j \geq k$ then we consider all possible $\binom{j}{k}$ possibilities to pick k edges out of the j edges. In this case, to reflect the choices for all possible $j \geq k$, we set $\Lambda_\alpha =: \Lambda_\alpha^k = \{(a_1, \ldots, a_k) \in \mathbb{N}^k : 1 \leq a_1 < a_2 < \cdots < a_k\}$, where the superscript emphasises that the edge in G has multiplicity k. If G has $e(G)$ multi-edges instead of just one multi-edge, with edge multiplicities $k_1, \ldots, k_{e(G)}$, and if $E(\alpha) = \{\alpha_1, \ldots, \alpha_{e(G)}\}$, then we let

$$\Lambda_\alpha = \Lambda_{\alpha_1}^{k_1} \otimes \Lambda_{\alpha_2}^{k_2} \otimes \cdots \otimes \Lambda_{\alpha_e(G)}^{k_{e(G)}} = \bigotimes_{i=1}^{e(G)} \Lambda_{\alpha_i}^{k_i}.$$

As an example, consider the multigraph $\mathcal{G}$ on three vertices in which there are two edges between each vertex pair. Then there are eight copies of the triangle graph G_1, but twelve copies of the multigraph G_2. Since $\rho(G_2)$=3, there are three isomorphic copies of G_2 in Γ, which we denote by $\alpha_1, \alpha_2, \alpha_3$. The indicator random variables are $X_{\alpha_i,(j)\otimes(k)\otimes(1,2)} = 1$, $i = 1, 2, 3$, $j = 1, 2$, and $k = 1, 2$. The total number of copies of the multigraph G is then defined as in (2.3).

Recall that in Section 2.1.1 we defined the random variable $Y_{a,b}$ to have probability mass function $\mathbb{P}(Y_{a,b} = k) = \pi_{a,b,k}$, $k = 0, 1, 2, \ldots$. Let $i_{(u,v)} \geq 1$ be the number of edges between vertices u and v in the fixed multigraph G. Then, by generalising the argument used to obtain

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

(2.5), we have

$$\begin{aligned}\mu(G) &= \sum_{a\in\Lambda_\alpha} \mathbb{E}X_{\alpha,a} \\ &= \sum_{c_1,c_2,\dots,c_{v(G)}=1}^{Q} f_{c_1} f_{c_2}\cdots f_{c_{v(G)}} \\ &\quad\times \prod_{1\le u<v\le v(G):\,(u,v)\in E(G)} \Bigg(\sum_{a_{(u,v)}=i_{(u,v)}}^{\infty} \binom{a_{(u,v)}-1}{i_{(u,v)}-1} \mathbb{P}(Y_{c_u,c_v}\ge a_{(u,v)})\Bigg).\end{aligned}$$

Here, to avoid double counting, we assume that the edges in the SBMM appear in some order, so picking the $a_{(u,v)}$th as well as $i_{(u,v)}-1$ smaller edges leads to the binomial coefficient in the above sum. Now,

$$\begin{aligned}\sum_{a=i}^{\infty}\binom{a-1}{i-1}\mathbb{P}(Y_{c_u,c_v}\ge a) &= \sum_{a=i}^{\infty}\binom{a-1}{i-1}\sum_{\ell=a}^{\infty}\mathbb{P}(Y_{c_u,c_v}=\ell) \\ &= \sum_{\ell=i}^{\infty}\sum_{a=i}^{\ell}\binom{a-1}{i-1}\mathbb{P}(Y_{c_u,c_v}=\ell) \\ &= \sum_{\ell=i}^{\infty}\binom{\ell}{i}\mathbb{P}(Y_{c_u,c_v}=\ell) \\ &= \mathbb{E}\left[\binom{Y_{c_u,c_v}}{i}\right],\end{aligned}$$

and so

$$\begin{aligned}\mu(G) &= \sum_{a\in\Lambda_\alpha} \mathbb{E}X_{\alpha,a} \\ &= \sum_{c_1,c_2,\dots,c_{v(G)}=1}^{Q} f_{c_1} f_{c_2}\cdots f_{c_{v(G)}} \prod_{1\le u<v\le v(G):\,(u,v)\in E(G)} \mathbb{E}\left[\binom{Y_{c_u,c_v}}{i_{(u,v)}}\right].\end{aligned} \tag{2.9}$$

(Note that when G is a simple graph, we have $i_{(u,v)}=1$, and so (2.9) reduces to (2.5).) The computation of $\mathbb{E}W$ is then as before, with (2.9) now replacing (2.5). Finally, we note that in the case of one class ($Q=1$) we can simplify (2.9). Suppose that G has $e_i(G)$ pairs of vertices which have i edges between them, with $e_{t(G)}(G)>0$ and $e_i(G)=0$ for all $i>t(G)$. Then

$$\mu(G) = \prod_{i=1}^{t(G)} \left(\mathbb{E}\left[\binom{Y_{1,1}}{i}\right]\right)^{e_i(G)}.$$

Remark 2.1. In [13], a Poisson approximation was obtained for the distribution of the number of isomorphic copies of a strictly balanced graph in the SBM. The possibility of multiple edges in the SBMM model means that, in general, a Poisson approximation will not be valid. This can be seen from the following simple example. Consider the SBMM(n, f, π) with edge probability distribution $\pi_{a,b,0}=1-p$, $\pi_{a,b,2}=p$ for all a, b, and $\pi_{a,b,k}=0$ for $k=1$ and $k\ge 3$ and all a, b. This model can be seen as the classical Erdős–Rényi random graph model with two

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

edges occurring (rather than just one) between two vertices independently and uniformly with probability p. Triangles (a member of class of strictly balanced graphs) occur in clumps of eight in this model, and the distribution of the total number of triangles is well-approximated by a $8\,\mathrm{Po}(\nu)$ distribution, where $\nu = \binom{n}{3}p^3$ is the expected number of triangles (this can be easily deduced from [10, Theorem 5.B]). Clearly, $8\,\mathrm{Po}(\nu)$ is not Poisson distributed. Instead, it is a special case of a compound Poisson distribution. A compound Poisson distribution is the distribution of the random sum $\sum_{n=1}^{N} X_i$, where the 'number of clumps' N follows a Poisson distribution and the 'clump sizes' X_i, $i = 1, 2, \ldots$ are independent and identically distributed (and can be constant, for example all equal to 8) and independent of N. Hence, in this paper we focus on compound Poisson approximations.

2.2. The Stein–Chen method for compound Poisson approximation

In this paper we use the Stein–Chen method for compound Poisson approximation to assess the distributional distance between $\mathcal{L}(W)$ and the compound Poisson $\mathrm{CP}(\boldsymbol{\lambda})$ distribution when the fixed graph G is a member of the class of strictly balanced graphs or strictly pseudo-balanced graphs. In this section we consider the case that G is a simple graph or multigraph in one framework. This distributional distance is measured using the total variation distance, which for nonnegative, integer-valued random variables U and V is

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

We begin by recalling the compound Poisson distribution, and then provide the relevant details of the Stein–Chen method for compound Poisson approximation.

The compound Poisson distributions $\mathrm{CP}(\boldsymbol{\lambda})$ is a family of distributions with an infinite-dimensional parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots)$ such that $\lambda_i \geq 0$ for all $i = 1, 2, \ldots$. We suppose that $\lambda := \sum_{i=1}^{\infty} \lambda_i < \infty$. Then a random variable having the $\mathrm{CP}(\boldsymbol{\lambda})$ distribution can be constructed as follows. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with probability mass function $\mathbb{P}(X_1 = i) = \lambda_i/\lambda$. Also, let Z be a Poisson distributed random variable independent of the $\{X_j\}$ with mean $\mathbb{E}Z = \lambda$. With these definitions of $\{X_j\}$ and Z, the random sum $\sum_{j=1}^{Z} X_j$ has the $\mathrm{CP}(\boldsymbol{\lambda})$ distribution. The $\mathrm{CP}(\boldsymbol{\lambda})$ distribution can be expressed in a second way. If Z_i are independent Poisson variables with means $\mathbb{E}Z_i = \lambda_i$ then $\sum_{i=1}^{\infty} i Z_i$ is $\mathrm{CP}(\boldsymbol{\lambda})$-distributed. In general, the probability mass function of a compound Poisson distribution is not available in closed form.

The Stein–Chen method for Poisson approximation was introduced by Chen [11], and was extended to compound Poisson approximation by Roos [27], [28]. A comprehensive account of the application of the Stein–Chen method for Poisson approximation in random graph theory can be found in [10], and the method was used to derive compound Poisson approximation of subgraph counts in Erdős–Rényi random graphs in [30]. Here, we present the compound Poisson framework given in [27], [28], and [30].

The random variable of interest is assumed to be of the form $W = \sum_{\alpha \in \Gamma} \sum_{a \in \Lambda_\alpha} X_{\alpha,a}$. The set $\Gamma \setminus \{\alpha\}$ is partitioned into three index classes for each $\alpha \in \Gamma$. These index classes are denoted by Γ_α^s, Γ_α^b, and Γ_α^w, so that $\Gamma \setminus \{\alpha\} = \Gamma_\alpha^s \cup \Gamma_\alpha^b \cup \Gamma_\alpha^w$. The intuition behind this partitioning is as follows. Think of W as a sum of random elements which are dependent, but the dependence is fairly local, with only a weak long-range dependence, if any at all. In this situation, one attempts to partition the index set such that the set Γ_α^s roughly consists of indices β such that the indicators $\{X_{\beta,b} \colon \beta \in \Gamma_\alpha^s\}$ are strongly dependent on $X_{\alpha,a}$ (for any $a \in \Lambda_a$ and $b \in \Lambda_\beta$); the set Γ_α^w roughly consists of indices β such that the indicators $\{X_{\beta,b} \colon \beta \in \Gamma_\alpha^w\}$ are very weakly dependent on $X_{\alpha,a}$; and the set Γ_α^b (boundary indices) roughly consists of indices β

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

such that the indicators $\{X_{\beta,b}\colon \beta \in \Gamma_\alpha^b\}$ are perhaps more than weakly dependent on $X_{\alpha,a}$. A natural choice for the sets Γ_α^s, Γ_α^w, and Γ_α^b in our situation is

$$\begin{aligned}\Gamma_\alpha^s &= \{\beta \in \Gamma \setminus \{\alpha\}\colon V(\alpha) = V(\beta)\},\\ \Gamma_\alpha^b &= \{\beta \in \Gamma\colon |V(\alpha) \cap V(\beta)| \in \{1, 2, \ldots, v(G) - 1\}\},\\ \Gamma_\alpha^w &= \{\beta \in \Gamma\colon |V(\alpha) \cap V(\beta)| = 0\}.\end{aligned}$$

For example, suppose that G is the path on three vertices and we consider the complete graph K_6 on six vertices. Ignoring redundant copies, on K_6 we have the $\binom{6}{3} \times 3 = 60$ copies of G in K_6, such as $\{(1,2),(2,3)\}$, $\{(1,3),(2,3)\}$, and $\{(1,2),(1,3)\}$. Suppose that $\alpha = \{\{1,2,3\},\{(1,2),(2,3)\}\}$. Then with our definition, $\Gamma_\alpha^s = \{\{(1,3),(2,3)\},\{(1,2),(1,3)\}\}$, while $\Gamma_\alpha^w = \{\{(4,5),(4,6)\},\{(4,5),(5,6)\},\{(4,6),(5,6)\}\}$, and $\Gamma_\alpha^b = \Gamma \setminus (\Gamma_\alpha^s \cup \Gamma_\alpha^w \cup \{\alpha\})$.

The set $\{\beta \in \Gamma_\alpha^s,\ b \in \Lambda_\beta\colon X_{\beta,b} = 1\}$ is the 'clump' about α. The size of the clump at α (which includes α itself) is denoted by Z_α and given by

$$Z_\alpha = \sum_{a \in \Lambda_\alpha} X_{\alpha,a} + \sum_{\beta \in \Gamma_\alpha^s} \sum_{b \in \Lambda_\beta} X_{\beta,b}.$$

With our choice of index sets, for multigraph counts $\{X_{\beta,b}\colon \beta \in \Gamma_\alpha^w\}$ is independent of $X_{\alpha,a}$.

The parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots)$ of the approximating compound Poisson distribution is

$$\lambda_i = \frac{1}{i} \sum_{\alpha \in \Gamma} \sum_{a \in \Lambda_a} \mathbb{E}[X_{\alpha,a}\, \mathbf{1}(Z_\alpha = i)], \qquad i = 1, 2, \ldots. \tag{2.10}$$

(This is a slight variant of the expression given in [27] and [28], which takes into account the presence of multiple edges.) The λ_i represent the expected number of 'clumps' of size i. They were calculated explicitly for many examples in [27], although none of these formulas apply to random graphs. Further details on the computation of the λ_i for subgraph counts in the SBMM can be found in Section 2.3.

Before stating the Stein–Chen bound for compound Poisson approximation that will be used in this paper, we introduce some further notation. Define

$$\eta_\alpha = \sum_{\beta \in \Gamma \setminus \Gamma_\alpha^w} \sum_{b \in \Lambda_\beta} X_{\beta,b}, \qquad V_\alpha = \sum_{\beta \in \Gamma_\alpha^b} \sum_{b \in \Lambda_\beta} X_{\beta,b},$$

$$\varepsilon = \sum_{\alpha \in \Gamma} \sum_{a \in \Lambda_\alpha} \big\{\mathbb{E}X_{\alpha,a}\mathbb{E}\eta_\alpha + \mathbb{E}[X_{\alpha,a} V_\alpha]\big\}.$$

Now we can state the Stein–Chen bound for Possion approximation (adapted from [28, Theorem 2]) that will be used in this paper:

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{CP}(\boldsymbol{\lambda})) \le c(\boldsymbol{\lambda})\varepsilon. \tag{2.11}$$

Here

$$c(\boldsymbol{\lambda}) = \sup_{A \subset \mathbb{Z}^+} \sup_{j \ge 0} |g_{\boldsymbol{\lambda},A}(j+1) - g_{\boldsymbol{\lambda},A}(j)|,$$

where $\mathbb{Z}^+ = \{0, 1, 2, \ldots\}$ and $g_{\boldsymbol{\lambda},A}(j)$ is the unique bounded solution of the so-called Stein equation

$$\sum_{i=1}^{\infty} i\lambda_i g(j+i) - jg(j) = \mathbf{1}(j \in A) - \mathbb{P}(Y \in A) \quad \text{for } Y \sim \mathrm{CP}(\boldsymbol{\lambda}).$$

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

It should be noted that our notation for η_α, which involves summing over $\beta \in \Gamma \setminus \Gamma_\alpha^w$ and, thus, includes α, is different from the analogous notation in [28], but is convenient in the subgraph counts applications considered in this paper.

There exist bounds in the literature for the constant $c(\boldsymbol{\lambda})$. Barbour *et al.* [9, Theorem 4] showed that

$$c(\boldsymbol{\lambda}) \le \mathrm{e}^{\lambda} \min\{1, \lambda_1^{-1}\} \quad \text{for all } \boldsymbol{\lambda} \tag{2.12}$$

with $\lambda = \sum_{i=1}^{\infty} \lambda_i < \infty$. In general, the dependence on $\boldsymbol{\lambda}$ in such a bound on $c(\boldsymbol{\lambda})$ cannot be improved, and the bound (2.12) is most useful when λ is small. However, in certain settings, the bound (2.12) can be improved; see [4], [7]–[9], and [14].

Remark 2.2. Roos [28, Theorem 2] contains an additional term which depends on

$$\phi_{\alpha,a,i} = \mathbb{E}|\mathbb{E}[X_{\alpha,a}\,\mathbf{1}(Z_\alpha = i) \mid (X_{\beta,b} \colon \beta \in \Gamma_\alpha^w, b \in \Lambda_\beta)] - \mathbb{E}[X_{\alpha,a}\,\mathbf{1}(Z_\alpha = i)]|.$$

By the independence of vertex-disjoint edges, and the fact that the edges in Γ_α^s are vertex-disjoint from those in Γ_α^w, we have Z_α is independent of the random variables $\{X_{\beta,b} \colon \beta \in \Gamma_\alpha^w, b \in \Lambda_\beta\}$. Therefore, $\phi_{\alpha,a,i} = 0$ for all values of α, a, and i. This means that this additional term vanishes from the bound of [28, Theorem 2].

In summary, bounding the total variation distance between the distribution of the subgraph counts in the SBMM and the CP($\boldsymbol{\lambda}$) distribution reduces to bounding the quantity ε. We will derive our compound Poisson approximations for subgraph counts (Theorems 3.1 and 4.1) using this approach.

2.3. Calculating the parameter of the limiting compound Poisson distribution

The majority of the approximation theorems derived in this paper will involve a compound Poisson approximation distribution of the subgraph counts in the SBMM, with the exception being the Poisson approximations of Section 5.2. In the case of Poisson approximation, the parameter $\nu := \mathbb{E}W$ of the limiting Po(ν) distribution is easily calculated using (2.7). For example, in the Erdős–Rényi $\mathfrak{G}(n, p)$ model, which is a special case of the SBMM, if the fixed graph G is a triangle then $v(G) = 3$, $\rho(G) = 1$, and $\mu(G) = p^3$, and so $\nu = \binom{n}{3} p^3$. In Section 5.2, it should be understood that $\nu = \mathbb{E}W$ is calculated via (2.7).

When a compound Poisson approximation is sought, the parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots)$ is given by (2.10). Given any fixed $\alpha^* \in \Gamma$, by symmetry we can write (2.10) as

$$\begin{aligned}
\lambda_i &= \frac{1}{i} \sum_{\alpha \in \Gamma} \sum_{a \in \Lambda_\alpha} \mathbb{E}X_{\alpha,a} \mathbb{P}(Z_\alpha = i \mid X_{\alpha,a} = 1) \\
&= \frac{1}{i} |\Gamma| \sum_{a \in \Lambda_{\alpha^*}} \mathbb{E}X_{\alpha^*,a} \mathbb{P}(Z_{\alpha^*} = i \mid X_{\alpha^*,a} = 1) \qquad (2.13) \\
&= \frac{1}{i} |\Gamma| \sum_{a \in \Lambda_{\alpha^*}} \mathbb{P}(Z_{\alpha^*} = i, X_{\alpha^*,a} = 1). \qquad (2.14)
\end{aligned}$$

In the simple graph case ($\pi_{a,b,i} = 0$ for $i \ge 2$ and all a, b), a further slight simplification of (2.13) is possible; see [30, Equation (20)].

For fixed graphs G in the SBMM, the conditional probability in (2.13) or the joint probability in (2.14) can be calculated on a case-by-case basis with the formulation (2.13) more suitable in a coupling context and (2.14) perhaps more suitable when local dependence is employed.

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

The conditional probability in (2.13) was calculated by [7] in the $\mathfrak{G}(n, p)$ setting for the Whisker graph (the graph on four vertices comprised of a vertex that is connected by an edge to one of the vertices of the complete graph on three vertices – it should be noted, though, that the Whisker graph is not strictly balanced and [7] used a different choice for the sets Γ^s_α, Γ^b_α, and Γ^w_α.) However, as noted by [30], it is not always possible to calculate the conditional probability exactly.

We now illustrate the computation of the λ_i via (2.14) for subgraphs in the SBMM. For the purposes of exposition, we consider the SBMM(n, π, f) model with one class, that is, $f_1 = 1$ and $f_i = 0$ for $i \geq 2$; the extension to multiple classes is not difficult, but involves more notation. Let $p_i = \pi_{1,1,i}$.

Example 2.1. We calculate $\boldsymbol{\lambda}$ for the case that G is the complete graph on v vertices with t edges between each vertex pair (a strictly pseudo-balanced graph). We have $|\Gamma| = \binom{n}{v}$, and for $\alpha^* \in \Gamma$, we have

$$\Lambda_{\alpha^*} = \Big\{(a_{1,1}, a_{2,1}, \dots, a_{t,1}) \otimes \cdots \otimes (a_{1,\binom{v}{2}}, a_{2,\binom{v}{2}}, \dots, a_{t,\binom{v}{2}}) \in \mathbb{N}^{t\binom{v}{2}} : \\ 1 \leq a_{1,1} < a_{2,1} < \cdots < a_{t,1}, \dots, 1 \leq a_{1,\binom{v}{2}} < a_{2,\binom{v}{2}} < \cdots < a_{t,\binom{v}{2}}\Big\}.$$

For $\alpha^* \in \Gamma$, we denote $a \in \Lambda_{\alpha^*}$ by $a = (a_{1,1}, \dots, a_{t,\binom{v}{2}})$. Also, $\Gamma^s_{\alpha^*} = \varnothing$ and so $Z_{\alpha^*} = \sum_{a \in \Lambda_{\alpha^*}} X_{\alpha^*,a}$. As G is a complete graph, clump sizes take the form of products of binomial coefficients only with one binomial coefficient per multi-edge. The joint probability is now

$$\mathbb{P}(Z_{\alpha^*} = i, X_{\alpha^*,a} = 1) = \sum_{x_1,\dots,x_{\binom{v}{2}} \in B_{a,v,t,i}} p_{x_1} \cdots p_{x_{\binom{v}{2}}}$$

for all $a = (a_{1,1}, \dots, a_{t,1}, \dots, a_{1,\binom{v}{2}}, \dots, a_{t,\binom{v}{2}}) \in A_{v,t,i}$, and is 0 otherwise. Here

$$A_{v,t,i} = \Bigg\{a_{1,1}, \dots, a_{t,\binom{v}{2}} : 1 \leq a_{1,1} < \cdots < a_{t,1}, \dots, 1 \leq a_{1,\binom{v}{2}} < \cdots < a_{t,\binom{v}{2}} \\ \text{and } 1 \leq \prod_{\ell=1}^{\binom{v}{2}} \binom{a_{t,\ell}}{t} \leq i\Bigg\},$$

$$B_{a,v,t,i} = \Bigg\{x_1, \dots, x_{\binom{v}{2}} : x_1 \geq a_{t,1}, \dots, x_{\binom{v}{2}} \geq a_{t,\binom{v}{2}} \text{ and } \prod_{\ell=1}^{\binom{v}{2}} \binom{x_\ell}{t} = i\Bigg\}.$$

Substituting into (2.14) then yields

$$\lambda_i = \frac{1}{i}\binom{n}{v} \sum_{a_{1,1},\dots,a_{t,\binom{v}{2}} \in A_{v,t,i}} \sum_{x_1,\dots,x_{\binom{v}{2}} \in B_{a,v,t,i}} p_{x_1} \cdots p_{x_{\binom{v}{2}}}, \qquad i \geq 1. \tag{2.15}$$

Note that in the case $t = 1$ and $p_i = 0$ for $i \geq 2$, the SBMM reduces to the $\mathfrak{G}(n, p_1)$ model and we have $\lambda_1 = \binom{n}{v} p_1^{\binom{v}{2}}$ and $\lambda_i = 0$ for $i \geq 2$. Thus, the limit distribution is simply the Poisson distribution $\mathrm{Po}(\binom{n}{v} p_1^{\binom{v}{2}})$, a result which is well known; see [10, Theorem 5.B].

For the edge graph ($v = 2$), (2.15) takes a particularly simple form, that is,

$$\lambda_{\binom{i}{t}} = \frac{1}{\binom{i}{t}}\binom{n}{2} \sum_{a_{1,1},\dots,a_{t,1} \in A_{2,t,\binom{i}{t}}} p_i = \frac{1}{\binom{i}{t}}\binom{n}{2} p_i \sum_{1 \leq a_{1,1} < a_{2,1} < \cdots < a_{t,1} \leq i} 1 = \binom{n}{2} p_i, \qquad i \geq t,$$

and $\lambda_k = 0$ for all other k.

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

Finally, we note that $\boldsymbol{\lambda}$ for the edge graph but now with Q classes is easily seen to be $\lambda_{\binom{i}{t}} = \binom{n}{2}\sum_{c_1,c_2=1}^{Q} f_{c_1} f_{c_2} \pi_{c_1,c_2,i}$, $i \geq t$, with $\lambda_k = 0$ for all other k. Equation (2.15) generalises to Q classes similarly.

3. Compound Poisson approximation in the SBMM: the case of simple fixed graphs

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

$$\alpha(G) = \min_{H} \frac{e(G) - e(H)}{v(G) - v(H)} \tag{3.1}$$

and, with (2.1),

$$\gamma(G) = \min_{H}\{d(G)v(H) - e(H)\} = \min_{H}[v(H)\{d(G) - d(H)\}], \tag{3.2}$$

where the minima are taken over all nonempty proper subgraphs $H \subsetneq G$. In interpreting the following theorem (see, for example, Remark 3.1), it is useful to note that if $\gamma(G) > 0$ or $\alpha(G) > d(G)$ then the graph G is strictly balanced; see [10]. Finally, we denote

$$\mu_1^* = \max_{a,b} \mathbb{E}Y_{a,b}, \tag{3.3}$$

where $Y_{a,b}$ is given in (2.6). With this notation, we can state our theorem.

Theorem 3.1. *Suppose that G is a strictly balanced graph. Then, with the notation of (2.2), (2.5), (2.10), and (3.1)–(3.3),*

$$\begin{aligned} &d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{CP}(\boldsymbol{\lambda})) \\ &\quad \leq M_{n,\pi,f}(G) \\ &\quad := \frac{c(\boldsymbol{\lambda})\rho(G)^2}{v(G)!} n^{v(G)} (\mu_1^*)^{e(G)} \left\{ \frac{v(G)^2}{v(G)!} n^{v(G)-1} (\mu_1^*)^{e(G)} \right. \\ &\qquad\qquad \left. + \sum_{i=1}^{v(G)-1} \binom{v(G)}{i} \frac{n^{v(G)-i} (\mu_1^*)^{\kappa(G,i)}}{(v(G)-i)!} \right\}, \end{aligned} \tag{3.4}$$

where $\kappa(G,i) = \max\{e(G) - id(G) + \gamma(G), (v(G) - i)\alpha(G)\}$. A bound for $c(\boldsymbol{\lambda})$ is given by (2.12).

Proof. We establish our bound by bounding $\varepsilon = \sum_{\alpha\in\Gamma}\sum_{a\in\Lambda_\alpha}\{\mathbb{E}X_{\alpha,a}\mathbb{E}\eta_\alpha + \mathbb{E}[X_{\alpha,a}V_\alpha]\}$. Combining (2.5) and (3.3) yields

$$\begin{aligned} \sum_{a\in\Lambda_\alpha} \mathbb{E}X_{\alpha,a} &= \sum_{c_1,c_2,\ldots,c_{v(G)}=1}^{Q} f_{c_1} f_{c_2} \cdots f_{c_{v(G)}} \prod_{1\leq u<v\leq v(G):\,(u,v)\in E(\alpha)} \mathbb{E}Y_{c_u,c_v} \\ &\leq \sum_{c_1,c_2,\ldots,c_{v(G)}=1}^{Q} f_{c_1} f_{c_2} \cdots f_{c_{v(G)}} (\mu_1^*)^{e(G)} \\ &= (\mu_1^*)^{e(G)}. \end{aligned} \tag{3.5}$$

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

Next we bound $\mathbb{E}\eta_\alpha = \sum_{\beta\in\Gamma\setminus\Gamma_\alpha^w}\sum_{b\in\Lambda_\beta}\mathbb{E}X_{\beta,b}$. Note that

$$\begin{aligned}
|\Gamma\setminus\Gamma_\alpha^w| &= |\{\beta\in\Gamma_\alpha^b\colon |V(\alpha)\cap V(\beta)|\geq 1\}| \\
&= \rho(G)\sum_{k=0}^{v(G)-1}\binom{n-v(G)}{k}\binom{v(G)}{v(G)-k} \\
&\leq \rho(G)v(G)\sum_{k=0}^{v(G)-1}\binom{n-v(G)}{n-v(G)-k}\binom{v(G)-1}{k} \\
&= \rho(G)v(G)\binom{(n-v(G))+(v(G)-1)}{v(G)-1} \\
&\leq \rho(G)v(G)\binom{n}{v(G)-1} \\
&\leq \rho(G)\frac{v(G)^2}{v(G)!}n^{v(G)-1},
\end{aligned}$$

where we use the fact that $\binom{n}{v(G)-1}\leq n^{v(G)-1}/v(G)!$. Combining this bound on $|\Gamma\setminus\Gamma_\alpha^w|$ with (3.5) yields

$$\mathbb{E}\eta_\alpha = \sum_{\beta\in\Gamma\setminus\Gamma_\alpha^w}\sum_{b\in\Lambda_\beta}\mathbb{E}X_{\beta,b}\leq |\Gamma\setminus\Gamma_\alpha^w|(\mu_1^*)^{e(G)}\leq \rho(G)\frac{v(G)^2}{v(G)!}n^{v(G)-1}(\mu_1^*)^{e(G)}. \tag{3.6}$$

From (3.5) and (3.6),

$$\begin{aligned}
\sum_{\alpha\in\Gamma}\sum_{a\in\Lambda_\alpha}\mathbb{E}X_{\alpha,a}\mathbb{E}\eta_\alpha &\leq \rho(G)\binom{n}{v(G)}(\mu_1^*)^{e(G)}\rho(G)\frac{v(G)^2}{v(G)!}n^{v(G)-1}(\mu_1^*)^{e(G)} \\
&\leq \rho(G)^2\frac{v(G)^2}{(v(G)!)^2}n^{2v(G)-1}(\mu_1^*)^{2e(G)}.
\end{aligned} \tag{3.7}$$

Finally, we bound the quantity $\sum_{\alpha\in\Gamma}\sum_{a\in\Lambda_\alpha}\mathbb{E}[X_{\alpha,a}V_\alpha]$. To do so, we partition Γ_α^b into sets $\{\Gamma_\alpha^{b,i}\}_{1\leq i\leq v(G)-1}$, where $\Gamma_\alpha^{b,i}=\{\beta\colon |\alpha\cap\beta|=i\}$. The cardinality of these sets can be bounded above by

$$|\Gamma_\alpha^{b,i}|\leq \rho(G)\binom{v(G)}{i}\binom{n}{v(G)-i}\leq \rho(G)\binom{v(G)}{i}\frac{n^{v(G)-i}}{(v(G)-i)!}. \tag{3.8}$$

We now bound the quantity $\sum_{a\in\Lambda_\alpha}\sum_{b\in\Lambda_\beta}\mathbb{E}[X_{\alpha,a}X_{\beta,b}]$, where $\beta\in\Gamma_\alpha^{b,i}$ for $1\leq i\leq v(G)-1$. It was shown in the proof of [13, Theorem 2.1] that for a strictly balanced graph G, isomorphic copies α and β that share i vertices have at least $e(G)+\kappa(G,i)$ edges in the union graph $\alpha\cup\beta$. Therefore, by following the steps used to obtain (2.5) and (3.5), we obtain

$$\sum_{a\in\Lambda_\alpha}\sum_{b\in\Lambda_\beta}\mathbb{E}[X_{\alpha,a}X_{\beta,b}]\leq (\mu_1^*)^{e(G)+\kappa(G,i)}\quad\text{for }\beta\in\Gamma_\alpha^{b,i}. \tag{3.9}$$

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

Combining (3.8) and (3.9) now yields

$$\begin{aligned}\sum_{\alpha\in\Gamma}\sum_{a\in\Lambda_\alpha}\mathbb{E}[X_{\alpha,a}V_\alpha] &= \sum_{\alpha\in\Gamma}\sum_{i=1}^{v(G)-1}\sum_{\beta_i\in\Gamma_\alpha^{b,i}}\sum_{a\in\Lambda_\alpha}\sum_{b_i\in\Lambda_{\beta_i}}\mathbb{E}[X_{\alpha,a}X_{\beta_i,b_i}]\\ &\le \rho(G)\binom{n}{v(G)}\sum_{i=1}^{v(G)-1}\rho(G)\binom{v(G)}{i}\frac{n^{v(G)-i}}{(v(G)-i)!}(\mu_1^*)^{e(G)+\kappa(G,i)}\\ &\le \rho(G)^2(\mu_1^*)^{e(G)}\frac{n^{v(G)}}{v(G)!}\sum_{i=1}^{v(G)-1}\binom{v(G)}{i}\frac{n^{v(G)-i}(\mu_1^*)^{\kappa(G,i)}}{(v(G)-i)!}. \qquad (3.10)\end{aligned}$$

Applying (2.11) together with the bounds (3.7) and (3.10) then yields (3.4). □

Remark 3.1. The result of Theorem 3.1 is perhaps most interesting when the limiting CP($\boldsymbol{\lambda}$) distribution is nondegenerate in the limit $n\to\infty$. For this to be the case, we require the mean $\mathbb{E}W$ to be nondegenerate, that is, $\mathbb{E}W$ does not tend to 0 or ∞ in the limit $n\to\infty$. Suppose that there exist universal constants c and C such that $cn^{-1/d(G)}\le\mathbb{E}Y_{a,b}\le Cn^{-1/d(G)}$ for all a, b. Then using the inequality

$$\frac{m^k}{k^k}\le\binom{m}{k}\le\frac{m^k}{k!},\qquad 1\le k\le m,$$

and (2.7), we obtain

$$\frac{\rho(G)}{v(G)^{v(G)}}c^{e(G)}\le\mathbb{E}W\le\frac{\rho(G)}{v(G)!}C^{e(G)}.$$

In this parameter regime,

$$\sum_{i=1}^{\infty}\lambda_i\le\sum_{i=1}^{\infty}i\lambda_i=\mathbb{E}W=O(1).$$

Therefore, from (2.12), we have $c(\boldsymbol{\lambda})\le K(G)$ for some $K(G)>0$ which does not involve n. Applying Theorem 3.1 then yields

$$d_{\mathrm{TV}}(\mathcal{L}(W),\mathrm{CP}(\boldsymbol{\lambda}))\le\frac{K(G)\rho(G)^2}{v(G)!}C^{e(G)}\left\{\frac{v(G)^2}{v(G)!}C^{e(G)}n^{-1}+\min(A,B)\right\}, \qquad (3.11)$$

where

$$A=(1+C^{\alpha(G)})^{v(G)-1}n^{1-\alpha(G)/d(G)},\qquad B=C^{e(G)+\gamma(G)}(1+C^{-d(G)})^{v(G)-1}n^{-\gamma(G)/d(G)}.$$

Example 3.1. Here we use (3.11) to obtain compound Poisson approximations for the number of copies of the following fixed graphs with $v\ge 3$ vertices in the SBMM(n,π,f) model. We consider the following strictly balanced graphs on v vertices each (the same choice as in [13, Remark 2.3]):

- $G_{1,v}$ represents a tree on the v vertices with $v-1$ edges;
- $G_{2,v}$ represents the cycle graph on the v vertices (with v edges);
- $G_{3,v}$ represents the complete graph on v vertices with one edge removed;
- $G_{4,v}$ represents the complete graph on v vertices.

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

TABLE 1: Values of $d(G)$, $\alpha(G)$, and $\gamma(G)$, with scaling and bounds on the rate of convergence.

G	$d(G)$	$\alpha(G)$	$\gamma(G)$	Scaling of μ_1^*	$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{CP}(\boldsymbol{\lambda}))$
$G_{1,v}$	$\frac{v-1}{v}$	1	$\frac{1}{v}$	$Cn^{-v/(v-1)}$	$O(n^{-1/(v-1)})$
$G_{2,v}$	1	$\frac{v-1}{v-2}$	1	Cn^{-1}	$O(n^{-1})$
$G_{3,v}$	$\frac{(v+1)(v-2)}{2v}$	$\frac{v^2-v-4}{2(v-2)}$	$\frac{1}{3}$ if $v=3$	$Cn^{-2v/(v+1)(v-2)}$	$O(n^{-1/2})$ if $v=3$
			1 if $v \geq 4$		$O(n^{-2/(v-1)})$ if $v \geq 4$
$G_{4,v}$	$\frac{v-1}{2}$	$\frac{v+1}{2}$	1	$Cn^{-2/(v-1)}$	$O(n^{-2/(v-1)})$

To apply (3.11), we must compute the quantities $d(G)$, $\alpha(G)$, and $\gamma(G)$ for each graph G. These values were computed in [13, Remark 2.3] and are presented in Table 1. If, for a given graph G, there exist universal constants c and C such that $cn^{-1/d(G)} \leq \mathbb{E}Y_{a,b} \leq Cn^{-1/d(G)}$ for all a, b (for which the limit distribution is nondegenerate), then a bound on $d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{CP}(\boldsymbol{\lambda}))$ now follows from (3.11). Table 1 also contains the appropriate scaling of $\mu_1^* = \max_{a,b} \mathbb{E}Y_{a,b}$ (note that all the $\mathbb{E}Y_{a,b}$ are all of the same order) as well as a bound on the rate of convergence in this parameter regime. For this rate of convergence, it is assumed that the proportion vector $f = f(n)$ remains constant as $n \to \infty$, and that G does not change with n. The conclusions we draw from Table 1 are very similar to those drawn in [13, Remark 2.3]; for example, 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.

We also note that if $\mathbb{E}Y_{a,b} = O(n^{-1})$ for all a, b (so that $\mu_1^* = O(n^{-1})$) then this gives rise to graphs with bounded average degree. With the scaling Cn^{-1} for μ_1^*, the number of isomorphic copies of $G_{2,v}$, the cycle graph on v vertices, is seen to have an approximation nondegenerate compound Poisson distribution. With the scaling $Cn^{-v/(v-1)}$ for μ_1^*, the number of copies of the tree graph $G_{1,v}$ also has an approximation nondegenerate compound Poisson distribution, and the average degree is $O(n^{-1/(v-1)})$; however, the scalings required to ensure that the limiting compound Poisson distribution for the number of copies of $G_{3,v}$ and $G_{4,v}$ is nondegenerate gives rises to graphs with unbounded average degree.

Example 3.2. The Poisson SBM from [19], for which the number of edges between vertices of type a and b follows a Poisson distribution with parameter $\omega_{a,b}$, falls into the framework of Theorem 3.1, with $\mu_1^* = \max_{a,b} \mathbb{E}Y_{a,b} = \max_{a,b} \omega_{a,b} = \omega^*$. According to Remark 3.1, the mean of the limiting compound distribution is nondegenerate if

$$cn^{-1/d(G)} \leq \omega_{a,b} \leq Cn^{-1/d(G)} \quad \text{for all } a, b.$$

In such a regime, bound (3.11) is applicable.

We can similarly apply Theorem 4.1 (below) to the Poisson SBM, although we omit the details. However, in Corollary 5.1 we calculate a Poisson approximation in the Poisson SBM.

We conclude this section by considering the setting of Theorem 3.1, but with edge probabilities (1.1) given by the more general

$$\mathbb{P}(\tilde{Y}_{i,j} = k \mid c_i = a, c_j = b) = \pi_{a,b,i,j,k}, \qquad k = 0, 1, 2, \ldots. \tag{3.12}$$

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

In this more general setting, the derivation of a bound on the total variation distance between $\mathcal{L}(W)$ and the limiting $\mathrm{CP}(\boldsymbol{\lambda})$ is almost unchanged. There is no change to the combinatorial arguments and the sole change is that we bound the expected number of edges by the quantity $\max_{a,b,i,j} \mathbb{E}Y_{a,b,i,j}$, where the random variable $Y_{a,b,i,j}$ has probability mass function

$$\mathbb{P}(Y_{a,b,i,j} = k) = \pi_{a,b,i,j,k}, \qquad k \geq 0.$$

In this way, we obtain the next corollary.

Corollary 3.1. *With the same setting as Theorem 3.1, but with edge probabilities given by (3.12) instead of (1.1), we have the bound*

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{CP}(\boldsymbol{\lambda})) \leq M^*_{n,\pi,f}(G), \tag{3.13}$$

where $M^*_{n,\pi,f}(G)$ is the bound (3.4) of Theorem 3.1 with the quantity $\mu^*_1 = \max_{a,b} Y_{a,b}$ replaced by the quantity $\max_{a,b,i,j} \mathbb{E}Y_{a,b,i,j}$.

Here $\boldsymbol{\lambda}$ is given by (2.10), but, in contrast to the setting of Theorem 3.1, we cannot compute $\boldsymbol{\lambda}$ using (2.14) because the edge probabilities (3.12) depend on the vertices at which the edge is incident. The results derived in the remainder of the paper generalise to edge probabilities of the form (3.12) similarly.

Example 3.3. Recall that the degree-corrected SBM of [19] has edge probabilities that follow the $\mathrm{Po}(\theta_1\theta_j\omega_{a,b})$ distribution. This model falls into the framework of Corollary 3.1 and we can apply bound (3.13) with $\max_{a,b,i,j} \mathbb{E}Y_{a,b,i,j} \leq (\theta^*)^2\omega^*$, where $\theta^* = \max_i \theta_i$ and $\omega^* = \max_{a,b} \omega_{a,b}$.

4. Compound Poisson approximation in the SBMM: extension to fixed multigraphs

In this section we generalise Theorem 3.1 to the case that the fixed graph G is a strictly pseudo-balanced multigraph. We begin by introducing appropriate generalisations of the notation in (3.1) and (3.2). With $f(G)$ denoting the number of pairs of vertices with at least one edge between them and $\partial(G)$ defined as in (2.8), let

$$\alpha_m(G) = \min_H \frac{f(G) - f(H)}{v(G) - v(H)}, \tag{4.1}$$

$$\gamma_m(G) = \min_H \{\partial(G)v(H) - e(H)\} = \min_H [v(H)\{\partial(G) - d(H)\}], \tag{4.2}$$

where the minima are taken over all nonempty proper subgraphs $H \subsetneq G$ such that $f(H) < f(G)$. In interpreting the following theorem, it is worth noting that the multigraph graph G is strictly pseudo-balanced if $\gamma_m(G) > 0$ or $\alpha_m(G) > \partial(G)$. Finally, let

$$\mu^*_k = \max_{a,b} \mathbb{E}[Y^k_{a,b}] \quad \text{and} \quad \mu^{**}_k = \max_{a,b} \mathbb{E}\left[\binom{Y_{a,b}}{k}\right]. \tag{4.3}$$

We are now ready to state our theorem.

Theorem 4.1. *Suppose that G is a strictly pseudo-balanced graph which has $e_i(G)$ pairs of vertices each of which have i edges between them, with $e_{t(G)}(G) > 0$ and $e_i(G) = 0$ for*

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

all $i > t(G)$. Then, with the notation (2.2), (2.9), (2.10), and (4.1)–(4.3),

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{CP}(\boldsymbol{\lambda})) \leq \frac{c(\boldsymbol{\lambda})\rho(G)^2}{v(G)!} n^{v(G)} \Bigg\{ \frac{v(G)^2}{v(G)!} n^{v(G)-1} \prod_{i=1}^{t(G)} ((\mu_i^{**})^{2e_i(G)}) + \sum_{i=1}^{v(G)-1} \binom{v(G)}{i} \frac{n^{v(G)-i} (\psi(G))^{e(G)+\kappa_m(G,i)}}{(v(G)-i)!} \Bigg\}, \tag{4.4}$$

where

$$\kappa_m(G,i) = \max\{e(G) - i\,\partial(G) + \gamma_m(G), (v(G)-i)\alpha_m(G)\},$$
$$\psi(G) = \max\Big\{2\mu^*_{2t(G)}, \max_{1\leq j\leq t(G)} (\mu_j^{**})\Big\}.$$

Proof. The proof follows along the same lines as that of Theorem 3.1. Again, we bound ε. First, combining (2.9) and (4.3) yields

$$\begin{aligned} \sum_{a\in\Lambda_\alpha} \mathbb{E}X_{\alpha,a} &= \sum_{c_1,c_2,\ldots,c_{v(G)}=1}^{Q} f_{c_1} f_{c_2} \cdots f_{c_{v(G)}} \prod_{1\leq u<v\leq v(G):\,(u,v)\in E(G)} \mathbb{E}\left[\binom{Y_{c_u,c_v}}{i_{(u,v)}}\right] \\ &\leq \sum_{c_1,c_2,\ldots,c_{v(G)}=1}^{Q} f_{c_1} f_{c_2} \cdots f_{c_{v(G)}} \prod_{i=1}^{t(G)} ((\mu_i^{**})^{e_i(G)}) \\ &= \prod_{i=1}^{t(G)} ((\mu_i^{**})^{e_i(G)}). \end{aligned} \tag{4.5}$$

We bound $\mathbb{E}\eta_\alpha$ in the same way as in the proof of Theorem 3.1 to obtain

$$\mathbb{E}\eta_\alpha \leq n^{v(G)-1}\rho(G)\frac{v(G)^2}{v(G)!}\prod_{i=1}^{t(G)} ((\mu_i^{**})^{e_i(G)}),$$

and, thus,

$$\sum_{\alpha\in\Gamma}\sum_{a\in\Lambda_\alpha} \mathbb{E}X_{\alpha,a}\mathbb{E}\eta_\alpha \leq \rho(G)^2 \frac{v(G)^2}{(v(G)!)^2} n^{2v(G)-1} \prod_{i=1}^{t(G)} ((\mu_i^{**})^{2e_i(G)}). \tag{4.6}$$

Finally, we bound the quantity $\sum_{\alpha\in\Gamma}\sum_{a\in\Lambda_\alpha}\mathbb{E}[X_{\alpha,a}V_\alpha]$. We can use bound (3.8) for the cardinality of the sets $\Gamma_\alpha^{b,i}$, $1 \leq i \leq v(G)-1$. However, we must adapt the argument given in Theorem 3.1 in order to bound the quantity

$$\sum_{a\in\Lambda_\alpha}\sum_{b\in\Lambda_\beta} \mathbb{E}[X_{\alpha,a}X_{\beta,b}], \tag{4.7}$$

where $\beta \in \Gamma_\alpha^{b,i}$ for $1 \leq i \leq v(G)-1$. Following the proof of Theorem 3.1, we note that for a strictly pseudo-balanced graph G, isomorphic copies α and β that share i vertices have at least $e(G) + \kappa_m(G,i)$ edges in the union graph $\alpha\cup\beta$. In bounding (4.7), we treat the following two cases separately: edge positions in which at least one vertex from the union graph $\alpha\cup\beta$ is not in $V(\alpha)\cap V(\beta)$, and edge positions that are in the intersection $V(\alpha)\cap V(\beta)$.

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

We first consider the case that at least one vertex from the union graph $\alpha \cup \beta$ is not in $V(\alpha) \cap V(\beta)$. Consider a pair of vertices (u, v) in which u has class a and v has class b. Suppose that at this edge position the number of edges in the multigraph G is j. Then by arguing as we did in establishing (2.9) and (4.5), we can bound the expected number occurrences of j edges at this edge position by μ_j^{**}.

Next, we consider the edges inside of the intersection $V(\alpha) \cap V(\beta)$. So, if α has $p > 0$ edges at (u, v) and β has $q > 0$ edges at (u, v), then the expected number of edges is

$$\sum_{k_1<k_2<\cdots<k_p,\, l_1<l_2<\cdots<l_q} \mathbb{P}(Y > k_p \vee l_q) = \sum_{k=p}^{\infty}\sum_{j=q}^{\infty}\binom{k-1}{p-1}\binom{j-1}{q-1}\mathbb{P}(Y \geq k \vee j), \quad (4.8)$$

where $x \vee y = \max\{x, y\}$ and we have set $Y = Y_{a,b}$ for the remainder of this proof. The above expression comes from considering the different ways in which we could choose the p edges we include in the subgraph located at α and the q in β. In this way we obtain a count for the number of ways in which we can choose the motif from the edges present once we have picked the location, as we may have more edges than necessary in one location leading to subgraphs appearing in each location with multiplicity. Also, we note that in bounding (4.8), without loss of generality, we can suppose that $p \geq q$. Using this argument, (4.8) can be bounded as

$$\begin{aligned}
&\sum_{k=p}^{\infty}\left\{\sum_{j=q}^{k}\binom{k-1}{p-1}\binom{j-1}{q-1}\mathbb{P}(Y \geq k) + \binom{k-1}{p-1}\sum_{j=k+1}^{\infty}\binom{j-1}{q-1}\mathbb{P}(Y \geq j)\right\} \\
&\quad = \sum_{k=p}^{\infty}\binom{k}{q}\binom{k-1}{p-1}\mathbb{P}(Y \geq k) + \sum_{j=p+1}^{\infty}\sum_{k=p}^{j-1}\binom{k-1}{p-1}\binom{j-1}{q-1}\mathbb{P}(Y \geq j) \\
&\quad = \sum_{k=p}^{\infty}\binom{k}{q}\binom{k-1}{p-1}\mathbb{P}(Y \geq k) + \sum_{j=p+1}^{\infty}\binom{j-1}{p}\binom{j-1}{q-1}\mathbb{P}(Y \geq j) \\
&\quad \leq \sum_{k=p}^{\infty}\binom{k}{q}\binom{k-1}{p-1}\mathbb{P}(Y \geq k) + \sum_{j=p+1}^{\infty}\binom{j}{p}\binom{j-1}{q-1}\mathbb{P}(Y \geq j),
\end{aligned}$$

where we use the fact that $\sum_{k=n}^{m}\binom{k}{n} = \binom{m+1}{n+1}$. Now we bound, for $k \geq p \geq q$,

$$\begin{aligned}
\binom{k}{q}\binom{k-1}{p-1} + \binom{k}{p}\binom{k-1}{q-1} &= \left(\frac{1}{q} + \frac{1}{p}\right)\frac{1}{(p-1)!\,(q-1)!}\frac{k!\,(k-1)!}{(k-q)!\,(k-p)!} \\
&\leq \frac{p+q}{p!\,q!}k^{p+q-1}.
\end{aligned}$$

Therefore, (4.8) can be bounded above by

$$\begin{aligned}
\frac{p+q}{p!\,q!}\sum_{k=0}^{\infty}k^{p+q-1}\mathbb{P}(Y \geq k) &\leq \frac{p+q}{p!\,q!}\sum_{k=0}^{\infty}k^{p+q}\mathbb{P}(Y = k) \\
&= \frac{p+q}{p!\,q!}\mathbb{E}[Y^{p+q}] \\
&\leq 2\mathbb{E}[Y^{p+q}] \\
&\leq 2\mathbb{E}[Y^{2t(G)}] \\
&\leq 2\mu_{2t(G)}^{*},
\end{aligned}$$

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

where the penultimate inequality holds since $\mathbb{E}Y \le \mathbb{E}[Y^2] \le \cdots \le \mathbb{E}[Y^{2t(G)}]$, and we used the fact that $(p+q)/p!\,q! \le 2$ for all $p, q \ge 1$. Therefore, for each edge within the intersection, the expected number of edges is bounded above by $2\mu^*_{2t(G)}$.

Combining these results, we obtain, for $\beta \in \Gamma_\alpha^{b,i}$,

$$\sum_{a\in\Lambda_\alpha}\sum_{b\in\Lambda_\beta}\mathbb{E}[X_{\alpha,a}X_{\beta,b}] \le \Big(\max\Big\{2\mu^*_{2t(G)}, \max_{1\le j\le t(G)}(\mu_j^{**})\Big\}\Big)^{e(G)+\kappa(G,i)} \le (\psi(G))^{e(G)+\kappa(G,i)}. \tag{4.9}$$

Combining (3.8) and (4.9) now yields

$$\begin{aligned}\sum_{\alpha\in\Gamma}\sum_{a\in\Lambda_\alpha}\mathbb{E}[X_{\alpha,a}V_\alpha] &= \sum_{\alpha\in\Gamma}\sum_{i=1}^{v(G)-1}\sum_{\beta_i\in\Gamma_\alpha^{b,i}}\sum_{a\in\Lambda_\alpha}\sum_{b_i\in\Lambda_{\beta_i}}\mathbb{E}[X_{\alpha,a}X_{\beta_i,b_i}]\\ &\le \rho(G)\binom{n}{v(G)}\sum_{i=1}^{v(G)-1}\rho(G)\binom{v(G)}{i}\frac{n^{v(G)-i}}{(v(G)-i)!}(\psi(G))^{e(G)+\kappa(G,i)}\\ &\le \rho(G)^2\frac{n^{v(G)}}{v(G)!}\sum_{i=1}^{v(G)-1}\binom{v(G)}{i}\frac{n^{v(G)-i}(\psi(G))^{e(G)+\kappa(G,i)}}{(v(G)-i)!}.\end{aligned} \tag{4.10}$$

Applying (2.11) together with the bounds (4.6) and (4.10) then yields (4.4). □

Remark 4.1. The limiting CP($\boldsymbol{\lambda}$) distribution is nondegenerate in the limit $n \to \infty$ if there exist universal constants c and C such that $cn^{-1/\partial(G)} \le \mathbb{E}[Y_{a,b}^k] \le Cn^{-1/\partial(G)}$ for all a, b and all $k = 1, 2, \ldots, 2t(G)$. In this case, arguing as we did in Remark 3.1 leads to

$$\frac{\rho(G)}{v(G)^{v(G)}}c^{e(G)} \le \mathbb{E}W \le \frac{\rho(G)}{v(G)!}C^{e(G)},$$

where $e(G)$ is the total number of edges in the multigraph G, and we also have $c(\boldsymbol{\lambda}) = O(1)$. There then exists a constant $K(G)$, not involving n, such that

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{CP}(\boldsymbol{\lambda})) \le K(G)\max\big\{n^{-1}, \min\{n^{1-\alpha_m(G)/\partial(G)}, n^{-\gamma_m(G)/\partial(G)}\}\big\}.$$

5. Further results

5.1. Extension to include self-loops

We consider the following generalisation of the SBMM(n, π, f) which includes multiple self-loops and thus allows for pseudo-graphs. The self-loop count random variables S_i, $i = 1, \ldots, n$, are assumed to be independent conditionally on the class of vertices, and the probability only depends on the class of the vertex, that is,

$$\mathbb{P}(\tilde{S}_i = k \mid i \in a) = \theta_{a,k}, \qquad k \ge 0.$$

In addition, we assume that $\tilde{S}_1, \ldots, \tilde{S}_n$ are mutually independent. We denote this model by SBMM$^*(n, \pi, \theta, f)$, where the '$*$' in the notation emphasises the generalisation of the SBMM model to include self-loops. When $\theta_{a,0} = 1$ for all $1 \le a \le Q$, the model reduces to the SBMM.

Let G be a fixed graph with possibly multiple edges and self-loops and $v(G) < n$ vertices. Denote by $V(G)$ and $E(G)$ its vertex and edge sets. The edge set will now include the additional

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

detail as to whether there are edge(s) from node i to node i, that is, whether there are self-loop(s) at vertex i. It is straightforward to generalise Theorems 3.1 and 4.1 to the SBMM$^*(n, \pi, \theta, f)$, since the inclusion of self-loops makes no difference to the combinatorial arguments of the proofs. We will only focus on the generalisation of Theorem 4.1; the generalisation of Theorem 3.1 follows from exactly the same argument.

The setup of Section 2.1.2 is mostly unchanged by the generalisation to self-loops, although the expectation (2.9) generalises to

$$\begin{aligned}\mu(G) &= \sum_{a\in\Gamma_\alpha} \mathbb{E}X_{\alpha,a} \\ &= \sum_{c_1,c_2,\dots,c_{v(G)}=1}^{Q} f_{c_1} f_{c_2} \cdots f_{c_{v(G)}} \\ &\quad \times \left\{ \prod_{1\le w\le V(G):\ (w,w)\in E(G)} \mathbb{E}S_{c_w} \right\} \left\{ \prod_{1\le u<v\le v(G):\ (u,v)\in E(G)} \mathbb{E}\left[\binom{Y_{c_u,c_v}}{i_{(u,v)}} \right] \right\}, \end{aligned} \tag{5.1}$$

where (w, w) denotes a self-loop at vertex w and the random variables S_a, $1 \le a \le Q$, have probability mass function $\mathbb{P}(S_a = k) = \theta_{a,k}$, $k = 0, 1, 2, \dots$. Using the convention that the empty product is set to 1, we recover (2.9) from (5.1) under the case of no self-loops. Definition (2.8) for strictly pseudo-balanced graphs is applicable to graphs G that have self-loops, and is equivalent to the condition that the subgraph G' formed by removing all self-loops from G is itself a strictly pseudo-balanced graph. The same is true of the notation $\alpha_m(G)$, $\gamma_m(G)$, and $\kappa_m(G, i)$ of Section 4. The analogue of (3.3) and (4.3) for self-loops is

$$\phi^* = \max_a \mathbb{E}S_a. \tag{5.2}$$

The proof of the following theorem is exactly the same as that of Theorem 4.1, except that the expectations $\mathbb{E}[X_{\alpha,a}X_{\beta,b}]$ are now alternated to take into account the self-loop probabilities, in the same way as in (5.1). For $\beta \in \Gamma_\alpha^{b,i}$, we compute the upper bound

$$\sum_{a\in\Lambda_\alpha} \sum_{b\in\Lambda_\beta} \mathbb{E}[X_{\alpha,a}X_{\beta,b}] \le (\phi^*)^{2s(G)-i} (\mu^*_{2t(G)})^{e(G)+\kappa(G,i)}.$$

The rest of the proof is unchanged and we obtain the following bound.

Theorem 5.1. *Suppose that G is a strictly pseudo-balanced graph, which has $e_i(G)$ pairs of vertices each of which have i edges between them, with $e_{t(G)}(G) > 0$ and $e_i = 0$ for all $i > t(G)$, and $s(G) \ge 0$ self-loops. Then, with the notation (2.2), (2.10), (4.1)–(4.3), (5.1), and (5.2),*

$$\begin{aligned} &d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{CP}(\boldsymbol{\lambda})) \\ &\quad \le \frac{c(\boldsymbol{\lambda})\rho(G)^2}{v(G)!} n^{v(G)} \left\{ \frac{v(G)^2}{v(G)!} n^{v(G)-1} (\phi^*)^{2s(G)} \prod_{i=1}^{t(G)} ((\mu_i^{**})^{2e_i(G)}) \right. \\ &\qquad \left. + \sum_{i=1}^{v(G)-1} \binom{v(G)}{i} \frac{n^{v(G)-i} (\phi^*)^{2s(G)-i} (\psi(G))^{e(G)+\kappa_m(G,i)}}{(v(G)-i)!} \right\}. \end{aligned}$$

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

5.2. Poisson approximation

Recalling Remark 2.1, in general a Poisson approximation for the distribution of the number of isomorphic copies of a strictly balanced graph is not valid in the SBMM(n, π, f) model. However, as we show in this section, in a certain parameter regime a Poisson approximation is valid. Of course, in the case of $\pi_{a,b,k} = 0$ for all $1 \le a, b \le Q$, and $k \ge 2$, the model reduces to the classical SBM and a Poisson approximation is valid in the small $\pi_{a,b,1}$ regime; see [13]. In Theorem 5.2 we obtain a parameter regime in which multiple edges are possible (that is, sufficient conditions on $\pi_{a,b,k} > 0$ for some $1 \le a, b \le Q$, and $k \ge 2$ such that a Poisson approximation is valid). Before stating the theorem, we present the Stein–Chen bound for Poisson approximation that we will use in the proof.

For $\alpha \in \Gamma$, let

$$A_\alpha = \{\beta \in \Gamma : |\alpha \cap \beta| \ge 1\},$$

so that $A_\alpha = \Gamma_\alpha^b \cup \Gamma_\alpha^s \cup \{\alpha\}$. Also, let

$$U_{\alpha,a} = \sum_{\beta \in A_\alpha} \sum_{b \in \Lambda_\beta} X_{\beta,b} - X_{\alpha,a}.$$

Then a simple corollary of [2, Theorem 1], or [10, Theorem 1.A], is that

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Po}(\nu)) \le \nu^{-1}(1 - \mathrm{e}^{-\nu}) \sum_{\alpha \in \Gamma} \sum_{a \in \Lambda_\alpha} \{\mathbb{E}X_{\alpha,a}\mathbb{E}\eta_\alpha + \mathbb{E}[X_{\alpha,a}U_{\alpha,a}]\}. \tag{5.3}$$

Here

$$\nu := \mathbb{E}W = \binom{n}{v(G)} \rho(G)\mu(G), \tag{5.4}$$

where we use (2.7) to write the expectation.

It is worth comparing (5.3) with the bound $d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{CP}(\boldsymbol{\lambda})) \le c(\boldsymbol{\lambda})\varepsilon$ that we used for compound Poisson approximation in Section 2.2. The term $\nu^{-1}(1-\mathrm{e}^{-\nu})$ is present since $c(\boldsymbol{\lambda}) = c((\nu, 0, 0, \ldots)) \le \nu^{-1}(1 - \mathrm{e}^{-\nu})$ for the $\mathrm{Po}(\nu)$ distribution; see [5, Lemma 4]. The quantity U_α becomes $U_{\alpha,a} = V_\alpha + T_{\alpha,a}$, where $T_{\alpha,a} = \sum_{\beta \in \Gamma_\alpha^s \cup \{\alpha\}} \sum_{b \in \Lambda_\beta} X_{\beta,b} - X_{\alpha,a}$. Therefore, (2.11) becomes

$$\begin{aligned} d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Po}(\lambda)) &\le \nu^{-1}(1 - \mathrm{e}^{-\nu}) \sum_{\alpha \in \Gamma} \sum_{a \in \Lambda_\alpha} \{\mathbb{E}X_{\alpha,a}\mathbb{E}\eta_\alpha + \mathbb{E}[X_{\alpha,a}V_\alpha]\} \\ &\quad + \nu^{-1}(1 - \mathrm{e}^{-\nu}) \sum_{\alpha \in \Gamma} \sum_{a \in \Lambda_\alpha} \mathbb{E}[X_{\alpha,a}T_{\alpha,a}]. \end{aligned} \tag{5.5}$$

We have already bounded the first term of (5.5) in Theorem 3.1, so to arrive at a Poisson approximation we must also be able to bound the second term.

Theorem 5.2. *Suppose that G is a strictly balanced graph. Then, with the notation (2.2), (2.5), (3.1)–(3.3), and (5.4),*

$$\begin{aligned} &d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Po}(\nu)) \\ &\quad \le M_{n,\pi,f}(G) + \nu^{-1}(1 - \mathrm{e}^{-\nu}) \frac{\rho(G)^2}{v(G)!} n^{v(G)} (\mu_1^*)^{e(G)-1} q_2^* \end{aligned}$$

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

$$= \nu^{-1}(1-\mathrm{e}^{-\nu})\frac{\rho(G)^2}{v(G)!}n^{v(G)}(\mu_1^*)^{e(G)-1}$$
$$\times\left\{\frac{v(G)^2}{v(G)!}n^{v(G)-1}(\mu_1^*)^{e(G)+1}+q_2^*+\sum_{i=1}^{v(G)-1}\binom{v(G)}{i}\frac{n^{v(G)-i}(\mu_1^*)^{\kappa(G,i)+1}}{(v(G)-i)!}\right\}, \quad (5.6)$$

where $M_{n,\pi,f}(G)$ is the bound (3.4) of Theorem 3.1 and $q_2^* = \max_{1\le a,b\le Q}\sum_{k=2}^{\infty}\pi_{a,b,k}$.

Proof. We bound $\sum_{\alpha\in\Gamma}\sum_{a\in\Lambda_\alpha}\mathbb{E}[X_{\alpha,a}T_{\alpha,a}]$, where $T_{\alpha,a} = \sum_{\beta\in\Gamma_\alpha^s\cup\{\alpha\}}\sum_{b\in\Lambda_\beta}X_{\beta,b} - X_{\alpha,a}$. First, we note that $|\Gamma_\alpha^s\cup\{\alpha\}| = \rho(G)$. To bound $\mathbb{E}[X_{\alpha,a}T_{\alpha,a}]$, we require bounds on $\sum_{a\in\Lambda_\alpha}\sum_{b\in\Lambda_\beta}\mathbb{E}[X_{\alpha,a}X_{\beta,b}]$, where $|\alpha\cap\beta| = v(G)$ and $(\alpha,a)\neq(\beta,b)$. For all such α,β,a,b, the largest value that $\sum_{a\in\Lambda_\alpha}\sum_{b\in\Lambda_\beta}\mathbb{E}[X_{\alpha,a}X_{\beta,b}]$ takes is when $\alpha=\beta$, since this minimises the number of edge positions in the intersection graph $\alpha\cap\beta$. There are $e(G)$ possible edge locations in the intersection graph, and since $(\alpha,a)\neq(\alpha,b)$, it follows that there must be at least two edges at one edge position and at least one edges at the remaining $e(G)-1$ edges positions. Hence, it follows that

$$\sum_{a\in\Lambda_\alpha}\sum_{b\in\Lambda_\beta}\mathbb{E}[X_{\alpha,a}X_{\beta,b}]\le q_2^*(\mu_1^*)^{e(G)-1}$$

for all α,β such that $|\alpha\cap\beta| = v(G)$. Therefore,

$$\sum_{\alpha\in\Gamma}\sum_{a\in\Lambda_\alpha}\mathbb{E}[X_{\alpha,a}T_{\alpha,a}]\le\rho(G)q_2^*(\mu_1^*)^{e(G)-1}.$$

Using this inequality to bound the second term of (5.5) and using bound (3.4) of Theorem 3.1 to bound the first term yields (5.6), as required. □

Remark 5.1. In the classic SBM(n,π,f) model, there are no multiple edges (that is, $\pi_{a,b,k}=0$ for all a,b if $k\ge2$), and, thus, $q_2^*=0$. Furthermore, $\mu_1^* = \max_{1\le a,b\le Q}\pi_{a,b,1} =: \pi^*$. Bound (5.6) then can be seen to be exactly the same as the bound of [13, Theorem 2.1], as one would expect.

Remark 5.2. In Remark 3.1, we showed that if there exist universal constants c and C such that $cn^{-1/d(G)}\le\mathbb{E}Y_{a,b}\le Cn^{-1/d(G)}$ for all x, then the limiting Po(ν) distribution is non-degenerate, and the bound $M_{n,\pi,f}(G)\to0$ as $n\to\infty$. If we also have $q_2^*\ll n^{-1/d(G)}$ then $d_{\mathrm{TV}}(\mathcal{L}(W),\mathrm{Po}(\nu))\to0$ as $n\to\infty$. Given that $\mathbb{E}Y_{a,b} = O(n^{-1/d(G)})$, the condition $q_2^*\ll n^{-1/d(G)}$ is equivalent to

$$q_2^* = \max_{1\le a,b\le Q}\left(\sum_{l=2}^{\infty}\pi_{a,b,l}\right)\ll\max_{1\le a,b\le Q}\pi_{a,b,1} = O(n^{-1/d(G)}). \quad (5.7)$$

A simple and natural example in which condition (5.7) is met is that of geometric probabilities $\pi_{a,b,k} = p^k(1-p)$, $k\ge0$, for all $1\le a,b\le Q$. The condition is met if $p\ll1$. The model of [19], which has Poisson probabilities $\pi_{a,b,k} = \mathrm{e}^{-\omega_{a,b}}\omega_{a,b}^k/k!$, $k\ge0$, also satisfies condition (5.7) if $\max_{1\le a,b\le Q}\omega_{a,b}\ll1$. Indeed, we obtain an explicit bound for the Poisson approximation of subgraph count distributions in this model in the following corollary.

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

Corollary 5.1. *Suppose we are in the setting of Theorem 5.2 and suppose that* $\pi_{a,b,k} = \mathrm{e}^{-\omega_{a,b}}\omega_{a,b}^k/k!$, $k \geq 0$. *Let* $\omega^* = \max_{1\leq a,b\leq Q}\omega_{a,b}$. *Then*

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Po}(\nu)) \leq \nu^{-1}(1-\mathrm{e}^{-\nu})\frac{\rho(G)^2}{v(G)!}n^{v(G)}(\omega^*)^{e(G)-1}\left\{\frac{v(G)^2}{v(G)!}n^{v(G)-1}(\omega^*)^{e(G)+1} + \frac{1}{2}(\omega^*)^2 + \sum_{i=1}^{v(G)-1}\binom{v(G)}{i}\frac{n^{v(G)-i}(\omega^*)^{\kappa(G,i)+1}}{(v(G)-i)!}\right\}. \quad (5.8)$$

Suppose now that there exist universal constants c *and* C *such that* $cn^{-1/d(G)} \leq \omega_{a,b} \leq Cn^{-1/d(G)}$ *for all* a, b. *Then*

$$\frac{\rho(G)}{v(G)^{v(G)}}c^{e(G)} \leq \nu \leq \frac{\rho(G)}{v(G)!}C^{e(G)}, \quad (5.9)$$

and so the limiting $\mathrm{Po}(\nu)$ *distribution is nondegenerate as* $n \to 0$. *Moreover, there then exists a constant* $K(G)$, *independent of* n, *such that*

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Po}(\nu)) \leq K(G)\max\left\{n^{-1}, n^{-1/d(G)}, \min\{n^{1-\alpha(G)/d(G)}, n^{-\gamma(G)/d(G)}\}\right\}. \quad (5.10)$$

Proof. We use bound (5.6) of Theorem 5.2, so we are required to bound μ_1^* and q_2^*. Recall from Example 3.2 that $\mu_1^* = \omega^*$. Also,

$$q_{a,b,2} = \sum_{k=2}^{\infty}\mathrm{e}^{-\omega_{a,b}}\frac{\omega_{a,b}^k}{k!} = 1-(1+\omega_{a,b})\mathrm{e}^{-\omega_{a,b}} \leq 1-(1+\omega^*)\mathrm{e}^{-\omega^*} \leq \frac{1}{2}(\omega^*)^2,$$

where the inequalities may be proved by simple calculus. Therefore, $q_2^* \leq \frac{1}{2}(\omega^*)^2$. Inserting these bounds into (5.6) yields (5.8), as required.

We obtain (5.9) by applying exactly the same argument as the one used in Remark 3.1. The bound (5.10) is also very similar to the total variation distance bound of Remark 3.1. Here we have an additional term to consider:

$$\frac{1}{2\nu}(1-\mathrm{e}^{-\nu})\frac{\rho(G)^2}{v(G)!}n^{v(G)}(\omega^*)^{e(G)+1}.$$

This term is of order $n^{-1/d(G)}$, and so we obtain (5.10), which completes the proof. □

Acknowledgements

M.C. acknowledges support from the Department of Statistics, University of Oxford, for a Summer Studentship. R.G. acknowledges support from the Engineering and Physical Sciences Research Council (EPSRC) (grant number EP/K032402/1) and is currently supported by a Dame Kathleen Ollerenshaw Research Fellowship. G.R. acknowledges support from the EPSRC (grant number EP/K032402/1) as well as from the Alan Turing Institute. The authors would like to thank the anonymous referees for their helpful comments and suggestions.

References

[1] Airoldi, E. M., Costa, T. B. and Chan, S. H. (2013). Stochastic blockmodel approximation of a graphon: theory and consistent estimation. In *Advances in Neural Information Processing Systems 26*.

[2] Arratia, R., Goldstein, L. and Gordon, L. (1989). Two moments suffice for Poisson approximations: the Chen–Stein method. *Ann. Prob.* **17,** 9–25.

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press

[3] Barbour, A. D. (1982). Poisson convergence and random graphs. *Math. Proc. Camb. Philos. Soc.* **92,** 349–359.
[4] Barbour, A. D. and Chryssaphinou, O. (2001). Compound Poisson approximation: a user's guide. *Ann. Appl. Prob.* **11,** 964–1002.
[5] Barbour, A. D. and Eagleson, G. K. (1983). Poisson approximation for some statistics based on exchangeable trials. *Adv. Appl. Prob.* **15,** 585–600.
[6] Barbour, A. D. and Röllin, A. (2017). Central limit theorems in the configuration model. Preprint. Available at https://arxiv.org/abs/1710.02644.
[7] Barbour, A. D. and Utev, S. (1998). Solving the Stein equation in compound Poisson approximation. *Adv. Appl. Prob.* **30,** 449–475.
[8] Barbour, A. D. and Xia, A. (2000). Estimating Stein's constants for compound Poisson approximation. *Bernoulli* **6,** 581–590.
[9] Barbour, A. D., Chen, L. H. Y. and Loh, W.-L. (1992). Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Prob.* **20,** 1843–1866.
[10] Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson Approximation*. Oxford University Press.
[11] Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Prob.* **3,** 534–545.
[12] Condon, A. and Karp, R. M. (1999). Algorithms for graph partitioning on the planted partition model. In *Randomization, Approximation, and Combinatorial Optimization*, Springer, Berlin, pp. 221–232.
[13] Coulson, M., Gaunt, R. E. and Reinert, G. (2016). Poisson approximation of subgraph counts in stochastic block models and a graphon model. *ESAIM Prob. Statist.* **20,** 131–142.
[14] Daly, F. (2017). On magic factors in Stein's method for compound Poisson approximation. *Electron. Commun. Prob.* **22,** 67.
[15] Daudin, J.-J., Picard, F. and Robin, S. (2008). A mixture model for random graphs. *Statist. Comput.* **18,** 173–183.
[16] Frank, O. and Strauss, D. (1986). Markov graphs. *J. Amer. Statist. Assoc.* **81,** 832–842.
[17] Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proc. Nat. Acad. Sci. USA* **99,** 7821–7826.
[18] Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983). Stochastic blockmodels: first steps. *Social Networks* **5,** 109–137.
[19] Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E(3)* **83,** 016107.
[20] Latouche, P. and Robin, S. (2016). Variational Bayes model averaging for graphon functions and motif frequencies inference in W-graph models. *Statist. Comput.* **26,** 1173–1185.
[21] Matias, C. and Robin, S. (2014). Modeling heterogeneity in random graphs through latent space models: a selective review. In *MMCS, Mathematical Modelling of Complex Systems*, EDP Sciences, Les Ulis, pp. 55–74.
[22] Milo, R. *et al.* (2002). Network motifs: simple building blocks of complex networks. *Science* **298,** 824–827.
[23] Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures Algorithms* **6,** 161–179.
[24] Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96,** 1077–1087.
[25] Olhede, S. C. and Wolfe, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proc. Nat. Acad. Sci. USA* **111,** 14722–14727.
[26] Picard, F. *et al.* (2008). Assessing the exceptionality of network motifs. *J. Comput. Biol.* **15,** 1–20.
[27] Roos, M. (1993). *Stein–Chen method for compound Poisson approximation.* Doctoral Thesis, University of Zürich.
[28] Roos, M. (1994). Stein's method for compound Poisson approximation: the local approach. *Ann. Appl. Prob.* **4,** 1177–1187.
[29] Sarajlić, A. *et al.* (2013). Network topology reveals key cardiovascular disease genes. *PLoS ONE* **8,** e71537.
[30] Stark, D. (2001). Compound Poisson approximation of subgraph counts in random graphs. *Random Structures Algorithms* **18,** 39–60.
[31] Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symp. Math. Statist. Prob.*, Vol. 2, University of California Press, Berkeley, pp. 583–602.
[32] Wegner, A. E. *et al.* (2018). Identifying networks with common organizational principles. To appear in *J. Complex Networks*. Available at https://doi.org/10.1093/comnet/cny003.

https://doi.org/10.1017/apr.2018.35 Published online by Cambridge University Press