Compound Poisson Approximations of Subgraph Counts in Random Graphs

Dudley Stark
Basic Research Institute in the Mathematical Sciences
HP Laboratories Bristol
HPL-BRIMS-98-22
October, 1998

Poisson approximation, random graphs, Stein's method

Poisson approximations for the counts of a given subgraph in large random graphs were accomplished using Stein's method by Barbour and others. Compound Poisson approximation results, on the other hand, have not appeared, at least partly because of the lack of a suitable coupling. We address that problem by introducing the concept of cluster determining pairs, leading to a useful coupling for a large class of subgraphs we call *local*. We find bounds on the compound Poisson approximation of counts of local subgraphs in large random graphs.

Compound Poisson approximations of subgraph counts in random graphs

Dudley Stark
BRIMS, Hewlett-Packard Laboratories
Filton Road, Stoke Gifford
Bristol BS12 6QZ, UK

October 15, 1998

Abstract

Poisson approximations for the counts of a given subgraph in large random graphs were accomplished using Stein's method by Barbour and others. Compound Poisson approximation results, on the other hand, have not appeared, at least partly because of the lack of a suitable coupling. We address that problem by introducing the concept of cluster determining pairs, leading to a useful coupling for a large class of subgraphs we call *local*. We find bounds on the compound Poisson approximation of counts of local subgraphs in large random graphs.

1 Introduction

Given a graph G, let $V(G)$ denote the vertex set of G and let $E(G)$ denote the edge set of G. We use the notation $v(G) = |V(G)|$ and $e(G) = |E(G)|$ to denote the number of vertices and edges in G. We use $\overline{G}$ to denote the complement of G. We write $F \subset H$ to mean F is a proper subgraph of H. Similarly, $E(F) \subset E(H)$ means $E(F)$ is a proper subset of $E(H)$. When two proper subgraphs F_1, F_2 of a graph G are isomorphic we write $F_1 \sim F_2$.

Let K_n be the complete graph on n vertices. The random graph $G(n,p)$ on n vertices is constructed by choosing each of the $\binom{n}{2}$ potential edges of

K_n to be present in $G(n,p)$ independently and with probability p. Fix a subgraph H of K_n. Let Γ be the set of $\alpha \subset K_n$ such that α is isomorphic to H and let $I_\alpha = \mathbf{1}(\alpha \subset G(n,p))$. The number of isomorphic copies of H in $G(n,p)$ is $W = \sum_\alpha I_\alpha$.

The distribution of the random variable W is of interest. The expectation of W is

$$\mathbb{E}W = \binom{n}{v(H)} p^{e(H)} \sim \frac{n^{v(H)} p^{e(H)}}{v(H)!}.$$

so that if p is chosen to be $p = cn^{-v(H)/e(H)}$, then $\mathbb{E}W \sim c^{e(H)}/e(H)!$. It is natural to study the distribution of W for these threshold values of p.

Let $d(H) = e(H)/v(H)$. Note that at threshold $np^{d(H)} = O(1)$. Given graph H, let $m(H) = \max\{F \subseteq H : d(F)\}$. A graph H is *balanced* if $d(H) = m(H)$. It is *strictly balanced* if $d(H) > m(H)$. It is *unbalanced* if $d(H) < m(H)$. It is well known (see [6] or [11]) that at threshold W converges weakly to a Poisson distribution if and only if H is strictly balanced. For unbalance graphs $W \Rightarrow \delta_0$. We require H to be balanced, but not strictly balanced, because we are interested in compound Poisson approximations of W. Moreover, in light of [7], we require that there exists a unique minimal subgraph $F \subset H$ such that $d(F) = d(H)$. We call F the *core* of H, written $\mathrm{core}(H)$.

The compound Poisson distribution $\mathrm{CP}(\boldsymbol{\lambda})$ has parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots)$ restricted to $\lambda := \sum_{i=1}^\infty \lambda_i < \infty$ and is distributed as $\sum_{i=1}^Z X_i$, where Z is Poisson with mean $\mathbb{E}Z = \lambda$, and the X_i are i.i.d. variables independent of Z with density $\mathbf{P}(X = i) = \lambda_i/\lambda$. Alternatively, $\mathrm{CP}(\boldsymbol{\lambda})$ is distributed as $\sum_{i=1}^\infty iZ_i$, where the Z_i are independent Poisson variables with means $\mathbb{E}Z_i = \lambda_i$.

Stein's method is a powerful technique for finding distributional approximations. The *Stein operator* $\mathcal{A}_{\boldsymbol{\lambda}}$ corresponding to $\mathrm{CP}(\boldsymbol{\lambda})$ is a functional on integer valued functions g defined by

$$(\mathcal{A}_{\boldsymbol{\lambda}}(g))(j) = jg(j) - \sum_{i=1}^\infty i\lambda_i g(j+i).$$

One shows that

$$|\mathbb{E}\mathcal{A}_{\boldsymbol{\lambda}}(g)| \leq \varepsilon_0 M_0(g) + \varepsilon_1 M_1(g) \tag{1}$$

for ε_0, ε_1 small, where M_0 and M_1 are given by $M_0(g) = \sup_{j\geq 1} |g(j)|$ and $M_1(g) = \sup_{j\geq 1} |g(j+1) - g(j)|$ for all bounded functions $g : \mathbb{N} \to \mathbb{R}$. If one

is using total variation distance then one takes $g = g_A$ to be the solution of the Stein equation

$$\mathcal{A}_{\boldsymbol{\lambda}}(g) = \mathbf{1}_A(j) - \mathrm{CP}(\boldsymbol{\lambda})\{A\},$$

in which case

$$d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{CP}(\boldsymbol{\lambda})) \leq \varepsilon_0 Q_0 + \varepsilon_1 Q_1,$$

where $Q_l = \sup_{A \subset \mathbb{Z}_+} M_l(g_A)$, $l = 0, 1$.

A general bound of the form (1) was given in [8]. For each α, define a set of indices $\Gamma_\alpha^{\mathrm{s}} \subset \Gamma \setminus \{\alpha\}$ for which I_α depends strongly on $(I_\beta : \beta \in \Gamma_\alpha^{\mathrm{s}})$. For each $\alpha \in \Gamma$, let $\mathcal{E}_\alpha$ be a random element and let χ be a variable taking on possible values of $\mathcal{E}_\alpha$. Suppose $\{I''_{\beta i\chi}(\alpha), \beta \in \Gamma\}$ and $\{I'_{\beta i\chi}(\alpha), \beta \in \Gamma\}$ are constructed for every i and χ on the same probability space in a way that

$$\mathcal{L}(I''_{\beta i\chi}(\alpha), \beta \in \Gamma) = \mathcal{L}(I_\beta, \beta \in \Gamma | I_\alpha I[Z_\alpha = i] = 1, \mathcal{E}_\alpha = \chi)$$

and

$$\mathcal{L}(I'_{\beta\chi}(\alpha), \beta \in \Gamma) = \mathcal{L}(I_\beta, \beta \in \Gamma).$$

Let $U_\alpha = \sum_{\beta \in \Gamma_\alpha^{\mathrm{s}}} I_\beta$, let $Z_\alpha = U_\alpha + I_\alpha$, and let $\theta_{\beta,\alpha,i}(b) = \mathbb{E}|I''_{\beta i\chi} - I'_{\beta i\chi}|$. Define the λ_i by

$$\lambda_i = \frac{1}{i} \sum_{\alpha \in \Gamma} \mathbb{E}\{I_\alpha I[Z_\alpha = i]\}.$$

Theorem 5 of [8] implies that (1) holds with

$$\varepsilon_0 = 0$$

and

$$\varepsilon_1 = \sum_{\alpha \in \Gamma} \left((\mathbb{E}I_\alpha)^2 + \mathbb{E}I_\alpha \mathbb{E}U_\alpha + \sum_{i=1}^{|\Gamma_\alpha^{\mathrm{s}}|+1} \sum_{\beta \in \Gamma_\alpha^{\mathrm{w}}} \mathbb{E}\{I_\alpha I[Z_\alpha = i] \theta_{\beta,\alpha,i}(\mathcal{E}_\alpha)\} \right). \quad (2)$$

A difficulty in the compound Poisson approximation of subgraph counts lies in finding appropriate couplings of $I'_{\beta i\chi}$ with $I''_{\beta i\chi}$. We will construct such couplings and use them to find compound Poisson approximations for counts of a subclass of graphs called 'local', which are defined in terms of the connectivity properties of $G(n,p)$ at the threshold for the existence of H.

We define cluster determining pairs in Section 2 and quantify compound Poisson approximations in terms of them. In Section 3, we define local graphs. We derive bounds on the compound Poisson approximation of subgraph counts in $G(n,p)$ for local graphs in Section 4.

2 Cluster determining pairs and local graphs

Fix $\alpha \sim H$, let $G'(n,p)$ be an independent copy of $G(n,p)$ and set $G'_\alpha(n,p) = G'(n,p) \cup \alpha$, where $\cup$ denotes the usual union of graphs. It is easy to check that $G'_\alpha(n,p) \stackrel{\mathrm{d}}{=} (G'(n,p)|I_\alpha = 1)$. This simple coupling suffices for bounding the Poisson approximation of subgraph counts of strictly balanced subgraphs in $G(n,p)$; see [3].

In our application to subgraph counts we will take $\Gamma^{\mathrm{s}}_\alpha = \{\beta \in \Gamma : \mathrm{core}(\beta) = \mathrm{core}(\alpha)\}$. Given a realization G of $G(n,p)$, we define the H-cluster at α to be the edge set defined by

$$\mathcal{C}_\alpha = \mathcal{C}_\alpha(G) = \bigcup_{\substack{\beta \in \Gamma^{\mathrm{s}}_\alpha \\ \beta \subset G}} E(\beta).$$

Note that $Z_\alpha(G(n,p)) = Z_\alpha(G'_\alpha(n,p))$ whenever $\mathcal{C}_\alpha(G(n,p)) = \mathcal{C}_\alpha(G'_\alpha(n,p))$, so that in using (2) it is useful to find some random element $\mathcal{E}_\alpha$ that determines $\mathcal{C}_\alpha$. A natural first approach might be to let $\mathcal{E}_\alpha = \mathcal{C}_\alpha(G'_\alpha(n,p))$ and construct $(I''_{\beta i\chi}, I'_{\beta\chi})$ by adding the edges of $\mathcal{E}_\alpha$ to $G(n,p)$ if they are not already there; we write $G(n,p) \cup \mathcal{E}_\alpha$ for the result. The problem with that idea is that it is possible, and indeed likely, that $\mathcal{C}_\alpha(G(n,p) \cup \mathcal{E}_\alpha) \neq \mathcal{C}_\alpha(G'_\alpha(n,p))$, hence $\mathcal{E}_\alpha$ does not determine $\mathcal{C}_\alpha$.

To define $\mathcal{E}_\alpha$ in a such a way that $\mathcal{E}_\alpha$ does determine $\mathcal{C}_\alpha$, we will choose $\mathcal{E}_\alpha = (E^+_\alpha, E^-_\alpha)$, where $E^+_\alpha \subset E(G'_\alpha(n,p))$ and $E^-_\alpha \subset E(\overline{G'_\alpha(n,p)})$ are edge sets. Thus, E^-_α is a set of edges that are *not* present in $G'_\alpha(n,p)$. Let $G(n,p) \star \mathcal{E}_\alpha$ be the graph obtained from $G(n,p)$ by adding all edges of E^+_α that are not present in $G(n,p)$, and by deleting all edges of E^-_α that are present in $G(n,p)$. If $\mathcal{E}_\alpha$ satisfies

$$\mathcal{C}_\alpha(G(n,p) \star \mathcal{E}_\alpha(G'_\alpha(n,p))) = \mathcal{C}_\alpha(G'_\alpha(n,p)), \tag{3}$$

then we will call $\mathcal{E}_\alpha$ a *cluster determining pair.*

Since cluster determining pairs determine Z_α, when we use them for compound Poisson approximation (2) may be simplified:

$$\varepsilon_1 \leq \sum_{\alpha \in \Gamma} \left((\mathbb{E}I_\alpha)^2 + \mathbb{E}I_\alpha \mathbb{E}U_\alpha + \sum_{\beta \in \Gamma^{\mathrm{w}}_\alpha} \mathbb{E}\{I_\alpha \theta_{\beta,\alpha}(\mathcal{E}_\alpha)\} \right), \tag{4}$$

where now $\{I''_{\beta\chi}(\alpha), \beta \in \Gamma\}$ and $\{I'_{\beta\chi}(\alpha), \beta \in \Gamma\}$ are constructed on the same probability space in a way that

$$\mathcal{L}(I''_{\beta\chi}(\alpha), \beta \in \Gamma) = \mathcal{L}(I_\beta, \beta \in \Gamma | \mathcal{E}_\alpha = \chi, I_\alpha = 1),$$

$$\mathcal{L}(I'_{\beta\chi}(\alpha), \beta \in \Gamma) = \mathcal{L}(I_\beta, \beta \in \Gamma),$$

and $\theta_{\beta,\alpha}(b) = \mathbb{E}|I''_{\beta\chi} - I'_{\beta\chi}|$. Moreover,

$$\varepsilon_1 \leq \sum_{\alpha\in\Gamma} \left((\mathbb{E}I_\alpha)^2 + \mathbb{E}I_\alpha \mathbb{E}U_\alpha + \mathbb{E}I_\alpha \mathbb{E}(W^+_\alpha + W^-_\alpha) \right), \tag{5}$$

where

$$W^+_\alpha = |\{\beta \in \Gamma^{\mathrm{w}}_\alpha : \beta \in G(n,p) \star \mathcal{E}_\alpha(G'_\alpha(n,p)) \text{ and } \beta \notin G(n,p)\}|$$

and

$$W^-_\alpha = |\{\beta \in \Gamma^{\mathrm{w}}_\alpha : \beta \subset G(n,p) \text{ and } \beta \notin G(n,p) \star \mathcal{E}_\alpha(G'_\alpha(n,p))\}|$$

are the numbers of copies of H added and deleted by performing the $\star$ operation.

How shall we define $\mathcal{E}_\alpha$? One possibility would be to choose $E^+ = E(G'(n,p))$ and $E^- = E(\overline{G'(n,p)})$. With this choice, $G(n,p) \star \mathcal{E}_\alpha = G'(n,p)$ and (3) is certainly satisfied. This choice is unsatisfactory, however, because $G'(n,p)$ is coupled independently with $G(n,p)$. We want to define $\mathcal{E}_\alpha$ in such a way that (3) is satisfied and such that the $\star$ operation does not change much of $G(n,p)$.

We call the graphs for which we have been able to use cluster determining pairs to get good compound Poisson approximations *local* graphs. To define local graphs, we need to introduce the concept of a *grading*; we use the definition of grading given in [5]. Suppose F' is a proper subgraph of F with F having k more vertices and l more edges than F'. We define the *additional degree* of F over F' as l/k. The *maximal additional degree* of F over F' is

$$m(F|F') = \max\{d(K|F') : F' \subseteq K \subseteq F, V(F') \neq V(F)\}.$$

We proceed as follows. Let H_0 be the unique minimal subgraph of H with $d(H_0) = m(H)$. Now, suppose we have inductively found subgraphs $H_0, H_2, \ldots, H_j$. If $H_j = H$, then we terminate the sequence. Otherwise, let

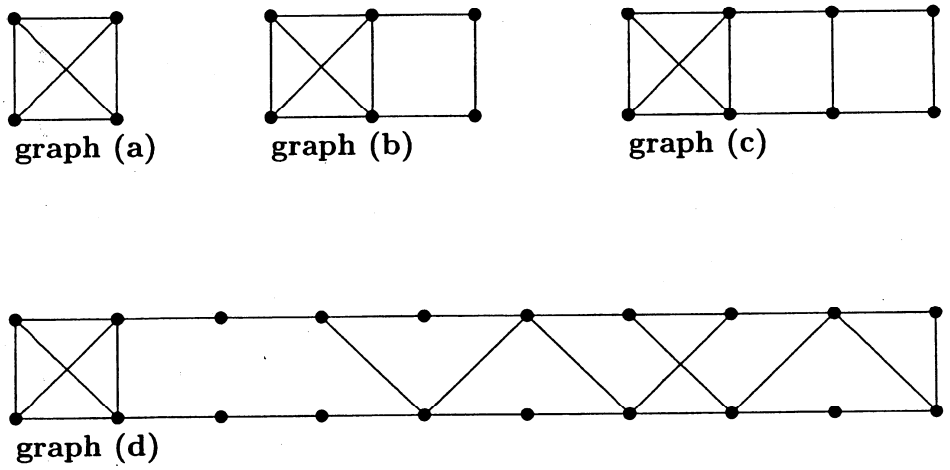

Figure 1: Graphs (a), (b), and (c) are local, but graph (d) is not.

H_{j+1} be a minimal subgraph of H with $V(H_j) \subset V(H_{j+1})$, $V(H_j) \neq V(H_{j+1})$, such that $d(H_{j+1}|H_j) = m(H|H_j)$. This sequence must end after finitely many terms, say at $H_k = H$. We call the sequence $(H_0, H_2, \ldots, H_k)$ a grading of H and call k the degree of the grading.

In Figure 1, graph (a) is strictly balanced and the unique grading has degree 0. The other graphs in Figure 1 are all balanced but not strictly balanced. Graph (b) has degree 2, graph (c) has degree 3, and, though it is not as obvious, graph (d) has degree 1. In all of the graphs, the core H_0 is a K_4 at the left of the graph. Graph (d) of Figure 1 has an unfortunate property, from our point of view. All of the graphs in Figure 1 have edge to vertex ratio 3 : 2. A path of length l connecting two vertices v and w is a series of edges $(v, v_1), (v_1, v_2), \ldots, (v_{l-1}, w)$. The extension of graph (d) contains paths of length of length 8 attached to the core. Let $p = n^{-2/3}$, the threshold for all of the graphs, and for graph (d) in particular. The threshold for having a path of length 8 connecting every vertex to the core of α occurs roughly when $n^7 p^8 = 2 \log n$ (see [5], Lemma 9, Section X.2) which is much less than $p = n^{-2/3}$. Paths of length 8 starting from the core of α penetrate the entire random graph, visiting every vertex. The number of isomorphic copies of H having the same core as α will be bounded a.s., so only $O(1)$ of these paths are contained in to any such extension. We must choose E_α^-

from the roughly n^2 edges of $\overline{G(n,p)}$. It seems like a difficult problem to find a cluster determining pair (E_α^+, E_α^-) for this graph in such a way that $\mathbb{E}|E_\alpha^-|$ is of order $o(n^2)$, say.

The compound Poisson approximation of W for Graph (b) is much more tractable than it is for graph (d). Let E_α^+ be all the edges of copies of graph (b) in $G'_\alpha(n,p)$ which have the same core as α. Let $\mathcal{V}_\alpha$ be the set of those vertices which are connected to a vertex of the core of α by an edge in $G'(n,p)$. Define E_α^- to be the edges in $\overline{G'_\alpha(n,p)}$ with at least one vertex in $\mathcal{V}_\alpha$, together with the edges in $\overline{G'_\alpha(n,p)}$ with a vertex in the core of α. Clearly, $\mathcal{C}_\alpha(G(n,p) \star \mathcal{E}_\alpha(G'_\alpha(n,p)))$ is at least as large as $\mathcal{C}_\alpha(G'_\alpha(n,p))$, by the way we have defined E_α^+. Note that, because of the way we have defined E_α^- any graph isomorphic to graph (b) and having the same core as α, but which is not contained in $G'(n,p)$, must have at least one of its edges in E_α^-. Therefore we have defined a cluster determining pair for graph (b). In addition, the number of vertices in $G'(n,p)$ connected by an edge of $G'(n,p)$ will be roughly $np = n^{1/3}$, and the number of edges in $\overline{G'(n,p)}$ with a vertex in the core of α is $O(n)$, so $|E_\alpha^-|$ is roughly $O(n^2p)+O(n)) = O(n^{4/3}) = o(n^2)$.

Graphs (a), (b), and (c) are all in the subclass of local graphs, which we will define in Section 3. Any graph with core K_4 made of repeated subextensions of the extension of graph (b) will also be local graphs. Graph (d), however, is not a local graph. We will be able to get compound Poisson approximations for local graphs at threshold and a little beyond. We will not be able to get compound Poisson approximations for graphs like graph (d), however. We know the number of copies of graph (d) in the random graph is asymptotically compound Poisson because of [7], but it is not yet clear how to use cluster determining pairs to show it. The class of local graphs is large and contains, for example, cycles with tree-like extensions.

We now proceed with the formal definition of local graphs. The distance $d_j(\mu, H_j)$ between a vertex $\mu \in V(H_{j+1})$ and H_j is defined to be the minimum path length of any path with edges lying entirely in H_{j+1}. (We know that such a path exists because of the uniqueness of H_0). We define the quantities ρ_j, $j = 0, \ldots, k-1$ to be $\rho_j = \max_{\mu \in H_{j+1}} d_j(\mu, H_j)$ and set $\rho = \max_{j \in [0,k-1]} \rho_j$. Let $J = \{i \in [0, k-1] : \rho_i = \rho\}$ and for each $i \in J$, let $\xi_i = \{v \in H_{i+1} : d_i(v, H_i) = \rho_i\}$. For each $i \in J$, we let Λ_i be the edges in H_i with both vertices in ξ_i. We call $\Lambda = \cup_i \Lambda_i$ is the set of *exterior edges.* All of the graphs in Figure 1 have exterior edges; for example in graph (b) there are two vertices in Λ_1 and an edge between them. The whisker graph in Figure

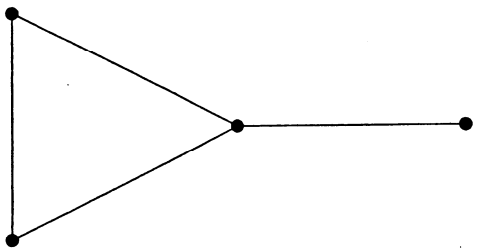

Figure 2: The whisker graph.

2 is a graph with no exterior edges. If $\Lambda \neq \emptyset$, set $\tau = \rho$; otherwise, set $\tau = \rho - 1$. A graph is called *local* there is a grading such that at threshold $(np)^\tau = o(n)$, or $n^{(1-v(H)/e(H))\tau} = o(n)$, or $(1 - v(H)/e(H))\tau < 1$. The relation $(np)^\tau = o(n)$ means that at threshold for the existence of H, the number of paths of length τ connecting any fixed pair of vertices tends to 0 a.s.

Which of the graphs in Figure 1 is not local? They all have $v(H)/e(H) = 2/3$. Graph (a) is strictly balanced, so it has a Poisson number of copies in $G(n,p)$. All of them have exterior edges, so $\tau = \rho$. Each graph has only one possible grading. Graph (b) has $k = 1$, $\rho_1 = 1$, $\tau = \rho = 1$, graph(c) has $k = 2$, $\rho_1 = 1$, $\rho_2 = 1$, $\tau = \rho = 1$, and graph(d) has $k = 1$, $\rho_1 = 5$, $\tau = \rho = 5 > 3$ so it is not local. For the whisker graph $k = 1$, $\rho_1 = 1$, $\tau = \rho - 1 = 0$ and it is local.

We now define the cluster determining pair that will be used. The set $E^+_\alpha = E^+_\alpha(G'_\alpha(n,p))$ is defined to be

$$E^+_\alpha = \bigcup_{\substack{\beta \in \Gamma^s_\alpha \\ \beta \subset G'_\alpha(n,p)}} E(\beta).$$

For each $i \in [1, k-1]$, let $\Gamma_{\alpha,i}$ be the set of subgraphs $\beta \subset K_n$ such that $\beta \sim H_i$, and $\text{core}(\beta) = \text{core}(\alpha)$. For each $i \in [0, k-1]$, let

$$\mathcal{C}_{\alpha,i} = \mathcal{C}_{\alpha,i}(G) = \bigcup_{\substack{\beta \in \Gamma_{\alpha,i} \\ \beta \subset G}} E(\beta).$$

Note that $\Gamma_{\alpha,k} = \Gamma^s_\alpha$ and $\mathcal{C}_{\alpha,k} = \mathcal{C}$. For each $i \in [0, k-1]$, let $\mathcal{V}_{\alpha,i}$ be the set of vertices connected to a vertex in $\mathcal{C}_{\alpha,i}$ by a path of length at most $\tau_i = \rho_i \wedge \tau$

lying completely in $G'_\alpha(n,p)$ and let $\mathcal{V}_\alpha$ be

$$\mathcal{V}_\alpha = \bigcup_{i=0}^{k-1} \mathcal{V}_{\alpha,i}.$$

We define E_α^- to be all edges in $\overline{G'_\alpha(n,p)}$ with at least one vertex in $\mathcal{V}_\alpha$.

We now show that $\mathcal{E}_\alpha$ is a cluster determining pair.

Lemma 1 *With this definition of* $\mathcal{E}_\alpha$,

$$\mathcal{C}_\alpha(G(n,p) \star \mathcal{E}_\alpha(G'_\alpha(n,p))) = \mathcal{C}_\alpha(G'_\alpha(n,p)).$$

Proof. We only consider subgraphs H with exterior edges; the argument for H without exterior edges is similar. It is obvious from the definition of E_α^+ that $\mathcal{C}_\alpha(G(n,p) \star \mathcal{E}_\alpha) \supseteq \mathcal{C}_\alpha(G'_\alpha(n,p))$. Let us suppose that

$$\mathcal{C}_\alpha(G(n,p) \star \mathcal{E}_\alpha(G'_\alpha(n,p)) \supset \mathcal{C}_\alpha(G'_\alpha(n,p)). \tag{6}$$

If (6) holds, then there is a least integer $0 < j^* \leq k$ such that there exists a $\beta \in \Gamma_{\alpha,j^*}$ satisfying $E(\beta) \subseteq \mathcal{C}_\alpha(G(n,p) \star \mathcal{E}_\alpha(G'(n,p)))$ and $E(\beta) \not\subseteq \mathcal{C}_\alpha(G'_\alpha(n,p))$. We will show that the existence of β leads to a contradiction.

Let γ be an isomorphic copy of H_{j^*-1} such that $\gamma \subset \beta$ and $\text{core}(\gamma) = \text{core}(\alpha)$. Define the set $E_{\beta,\gamma} = E(\beta) \setminus E(\gamma)$. Assume first that $E_{\beta,\gamma} \cap E_\alpha^- = \emptyset$. All of the edges of $E_{\beta,\gamma}$ with a vertex in $V(\gamma)$ must then be in $G'_\alpha(n,p)$, or else they would be in E_α^-. By induction, all of the edges of $E_{\beta,\gamma}$ with a vertex connected to a vertex in $V(\gamma)$ by a path in β of length r, $r \leq \rho_{j^*-1}$, must be in $G'_\alpha(n,p)$. But then β is a subgraph of $G'_\alpha(n,p)$ and $E(\beta) \subseteq \mathcal{C}_\alpha(G'_\alpha(n,p))$, contradicting our choice of β.

If we assume that $E_{\beta,\gamma} \cap E_\alpha^- \neq \emptyset$, then there is some $\mu \in E(\beta)$ such that $\mu \in E_\alpha^-$. But then $\mu \notin \mathcal{C}_\alpha(G(n,p) \star \mathcal{E}_\alpha(G'(n,p)))$. Hence, we have $E(\beta) \not\subseteq \mathcal{C}_\alpha(G(n,p) \star \mathcal{E}_\alpha(G'(n,p)))$, once again contradicting our choice of β. ■

3 Compound Poisson approximations

Define

$$\kappa(H) = \min_F \left(\frac{e(H)}{v(H)} v(F) - e(F) \right), \tag{7}$$

where the minimum is taken over all subgraphs $F \subset H$ such that $d(F) < d(H)$. We know $\kappa(H) > 0$ because H is balanced. Define $\Psi = \Psi(n,p)$ to be

$$\Psi = \max_{0 \le i \le k-1} n^{v(H_i)-v(H_0)} p^{e(H_i)-e(H_0)} = O\left((\mathbb{E}W)^{(v(H_{k-1})-v(H_0))/v(H)} \vee 1 \right),$$

and define $\Phi = \Phi(n,p)$ by

$$\Phi = \max_{0 \le i \le k-1} n^{v(H_i)} p^{e(H_i)} = O\left((\mathbb{E}W)^{v(H_{k-1})/v(H)} \vee 1 \right).$$

We let $C > 0$ represent some constant which may change from line to line.

Proposition 1 *For some constant $C > 0$ depending only on H,*

$$\begin{aligned} \varepsilon_1 \le\ & C\Big\{ \mathbb{E}W p^{e(H)} + (\mathbb{E}W)^2 n^{-v(H_0)} + \mathbb{E}W\,(\mathbb{E}W \vee 1)\, n^{-\kappa(H)/d(H)} \\ & + (\mathbb{E}W)^2 \Phi\, n^{-\kappa(H)/d(H)} \mathbf{1}(\tau \neq 0) + (\mathbb{E}W)^2 \Psi\, \frac{(np \vee 1)^\tau}{n} \\ & + (\mathbb{E}W)^2 \Psi \left(\frac{(np)^\tau}{n} \vee p \right) \mathbf{1}(\tau \neq 0) \Big\}. \end{aligned}$$

Proof. We will bound each of the terms on the right hand side of (5). First, note that

$$\sum_{\alpha \in \Gamma} (\mathbb{E}I_\alpha)^2 = O\left(n^{v(H)} p^{2e(H)} \right) = O\left(\mathbb{E}W p^{e(H)} \right) \tag{8}$$

and

$$\sum_{\alpha \in \Gamma} \mathbb{E}I_\alpha \mathbb{E}U_\alpha = O\left(n^{v(H)} p^{e(H)} n^{v(H)-v(H_0)} p^{e(H)} \right) = O\left((\mathbb{E}W)^2 n^{-v(H_0)} \right). \tag{9}$$

The next task is to bound $\mathbb{E}W_\alpha^-$. By definition, each edge in E_α^- has a vertex in $\mathcal{V}_\alpha$. Therefore, W_α^- is bounded by the number of copies of H in $G(n,p)$ with vertices in $\mathcal{V}_\alpha$. We have

$$\begin{aligned} \mathbb{E}W_\alpha^- &= \mathbb{E}\left(\mathbb{E}(W_\alpha^- | \mathcal{V}_\alpha) \right) \\ &\le \mathbb{E}\left(v(H) \mathcal{V}_\alpha n^{v(H)-1} p^{e(H)} \right) \\ &= O\left(\mathbb{E}W\, \mathbb{E}\mathcal{V}_\alpha n^{-1} \right). \end{aligned} \tag{10}$$

The factor $v(H)\mathcal{V}_\alpha n^{v(H)-1}$ at (10) is a crude upper bound on the number of ways of choosing $v(H)$ vertices, at least one of which is contained in $\mathcal{V}_\alpha$. Consider the graphs $\delta_i \cup \gamma$, where $\delta_i \in \Gamma_{\alpha,i}$ and γ is a path of length at most τ_i with at least one vertex in δ_i. We clearly have

$$\mathcal{V}_\alpha \le \sum_{i=0}^{k-1} \sum_{\delta_i \in \Gamma_{\alpha,i}} \sum_{\gamma} I[\delta_i \cup \gamma \subset G'_\alpha(n,p)].$$

Therefore,

$$\begin{aligned} \mathbb{E}\mathcal{V}_\alpha &= \sum_{i=0}^{k-1} O\left(n^{v(H_i)-v(H_0)} p^{e(H_i)-e(H_0)} (np \vee 1)^{\tau_i}\right) \\ &= O\left(\Psi\,(np \vee 1)^{\tau}\right). \end{aligned}$$

Hence,

$$\mathbb{E}W_\alpha^- = O\left(\mathbb{E}W\ \Psi \frac{(np \vee 1)^{\tau}}{n}\right). \tag{11}$$

Next we bound $\mathbb{E}W_\alpha^+$. The edges in $G(n,p) \star \mathcal{E}_\alpha(G'_\alpha(n,p))$ are a subset of the edges in $G(n,p) \cup E_\alpha^+(G'_\alpha(n,p))$, so that

$$\begin{aligned} W_\alpha^+ &\le \sum_{\beta \in \Gamma_\alpha^w} I[\beta \subset G(n,p) \star \mathcal{E}_\alpha(G'_\alpha(n,p))]\, I[\beta \not\subset G(n,p)] \\ &\le \sum_{\beta \in \Gamma_\alpha^w} I[\beta \subset G(n,p) \cup E_\alpha^+(G'_\alpha(n,p))]\, I[\beta \not\subset G(n,p)] \\ &\le \sum_{\substack{\beta \in \Gamma_\alpha^w \\ V(\beta) \cap V(\alpha) \neq \emptyset}} I[\beta \subset G(n,p) \cup E_\alpha^+(G'_\alpha(n,p))] \qquad (12) \\ &\quad + \sum_{\substack{\beta \in \Gamma_\alpha^w \\ V(\beta) \cap V(\alpha) = \emptyset}} I[\beta \subset G(n,p) \cup E_\alpha^+(G'_\alpha(n,p))]\, I[\beta \not\subset G(n,p)]. \qquad (13) \end{aligned}$$

The calculation $1 - (1-p)^2 = p(2-p)$ shows that

$$G(n,p) \cup G'_\alpha(n,p) \stackrel{\mathrm{d}}{=} G_\alpha(n, p(2-p)).$$

Therefore, the expectation of (12) is

$$\sum_{\substack{\beta \in \Gamma_\alpha^w \\ V(\beta) \cap V(\alpha) \neq \emptyset}} \mathbf{P}\left(\beta \subset G(n,p) \cup E_\alpha^+(G'_\alpha(n,p))\right)$$

$$\leq \sum_{\substack{\beta\in\Gamma_\alpha^{\mathrm{w}}\\ V(\beta)\cap V(\alpha)\neq\emptyset}} \mathbf{P}(\beta\subset G_\alpha(n,p(2-p))). \tag{14}$$

Since $\beta\in\Gamma_\alpha^{\mathrm{w}}$, by definition $\mathrm{core}(\beta)\neq\mathrm{core}(\alpha)$, which implies that $\beta\cap\alpha\sim F$ for some subgraph $F\subset H$ such that $d(F)<d(H)$. Therefore,

$$\begin{aligned}
\text{Expression}(14) &\leq C\sum_{F:d(F)<d(H)} n^{v(H)-v(F)}(p(2-p))^{e(H)-p(F)}\\
&\leq C\sum_{F:d(F)<d(H)} (\mathbb{E}W)^{1-e(F)/e(H)}n^{-(v(F)-v(H)e(F)/e(H))}\\
&\leq \begin{cases} C\,(\mathbb{E}W)\,n^{-\kappa(H)/d(H)} & \text{if } \mathbb{E}W\geq 1;\\ Cn^{-\kappa(H)/d(H)} & \text{if } \mathbb{E}W<1.\end{cases}\\
&\leq C\,(\mathbb{E}W\vee 1)\,n^{-\kappa(H)/d(H)}.
\end{aligned}\tag{15}$$

This part of the argument has followed part of the proof of Theorem 5.B of [3].

We now break (13) into two terms. For each $\beta\in\Gamma_\alpha^{\mathrm{w}}$ such that $V(\beta)\cap V(\alpha)=\emptyset$, there is either an i such that there exists $\delta_i\in\Gamma_{\alpha,i}$ such that $\delta_i\subset G(n,p)\star\mathcal{E}_\alpha(G'_\alpha(n,p))$ and $V(\delta_i)\cap V(\beta)\neq\emptyset$, or else $V(\delta_i)\cap V(\beta)=\emptyset$ for all δ_i such that $\delta_i\subset G(n,p)\star\mathcal{E}_\alpha(G'_\alpha(n,p))$. For all $i\in[0,k-1]$ and each $\delta_i\in\Gamma_{\alpha,i}$ such that $V(\delta_i)\cap V(\beta)=\emptyset$, we consider all paths of length at most τ_i connecting a vertex of δ_i with a vertex in β. We write $\Gamma(\delta_i,\beta)$ for the set of such paths. For each $\beta\in\Gamma_\alpha^{\mathrm{s}}$, $\beta\subset G(n,p)\star E_\alpha^+(G'_\alpha(n,p))$ such that $V(\delta_i)\cap V(\beta)=\emptyset$ for all i and each $\delta_i\in\Gamma_{\alpha,i}$, there must be at least one δ_i such that there exists a path $\gamma\in\Gamma(\delta_i,\beta)$ such that $\delta_i\cup\gamma\cup\beta\subset G(n,p)\star E_\alpha^+(G'_\alpha(n,p))$. Therefore, (13) is bounded by

$$\sum_{\substack{\beta\in\Gamma_\alpha^{\mathrm{w}}\\ V(\beta)\cap V(\alpha)=\emptyset}}\sum_{i=0}^{k-1}\sum_{\substack{\delta_i\in\Gamma_{\alpha,i}\\ V(\delta_i)\cap V(\beta)\neq\emptyset}} I[\delta_i\cup\beta\subset G(n,p)\cup E_\alpha^+(G'_\alpha(n,p))] \tag{16}$$

$$+\sum_{\substack{\beta\in\Gamma_\alpha^{\mathrm{w}}\\ V(\beta)\cap V(\alpha)=\emptyset}}\sum_{i=0}^{k-1}\sum_{\substack{\delta_i\in\Gamma_{\alpha,i}\\ V(\delta_i)\cap V(\beta)=\emptyset}}\sum_{\gamma\in\Gamma(\delta_i,\beta)} I[\delta_i\cup\gamma\cup\beta\subset G(n,p)\cup E_\alpha^+(G'_\alpha(n,p))]. \tag{17}$$

The expectation of (16) is bounded by

$$\sum_{\substack{\beta\in\Gamma_\alpha^{\mathbf{w}}\\ V(\beta)\cap V(\alpha)=\emptyset}}\sum_{i=0}^{k-1}\sum_{\substack{\delta_i\in\Gamma_{\alpha,i}\\ V(\delta_i)\cap V(\beta)\neq\emptyset}}\mathbf{P}\left(\delta_i\cup\beta\subset G_\alpha(n,p(2-p))\right). \tag{18}$$

If $\tau=0$, then (18) is an empty sum, because in that case $\delta_0=\mathrm{core}(\alpha)$. Assume now that $\tau>0$. In bounding (18) we first choose β in $O\left(n^{v(H)}\right)$ ways, each of which has probability $(p(2-p))^{e(H)}$ of appearing in $G_\alpha(n,p(2-p))$. Thus, for some constant $C>0$, (18) is bounded by

$$C\left(n^{v(H)}p^{e(H)}\right)\sup_{\substack{\beta\in\Gamma_\alpha^{\mathbf{s}}\\ V(\alpha)\cap V(\beta)=\emptyset}}\sup_i\sup_{\substack{\delta_i\in\Gamma_{\alpha,i}\\ V(\delta_i)\cap V(\beta)\neq\emptyset}}n^{v(\delta_i)-v(\delta_i\cap(\alpha\cup\beta))}p^{e(\delta_i)-e(\delta_i\cap(\alpha\cup\beta))}.$$

Because we are assuming that δ_i intersects α and β that are vertex disjoint, because δ_i is a subgraph of $\alpha\sim H$, and because H has a unique minimal subgraphs H_0 such that $d(H_0)=d(H)$, we conclude that $d(\delta_i\cap(\alpha\cup\beta))<d(H)$. Thus, (18) is bounded by

$$C(\mathbb{E}W)\max_{0\le i\le k-1}\max_{F_i}n^{v(H_i)-v(F_i)}p^{e(H_i)-e(F_i)},$$

where the max is taken over subgraphs $F_i\subset H_i$ such that $d(F_i)<d(H_i)$. By arguing as for (15), we get the bound

$$\begin{aligned}\text{Expression}(18) &\le C(\mathbb{E}W)\left(\max_{0\le i\le k-1}n^{v(H_i)}p^{e(H_i)}\right)n^{-\kappa^*/d^*}\\ &= C(\mathbb{E}W)\Phi n^{-\kappa^*/d^*},\end{aligned}$$

where d^* is given by

$$d^*=\max_i d(H_i)=\max_i d(H)=d(H)$$

and κ^* is given by

$$\begin{aligned}\kappa^* &= \min_i\min_{F_i}\left(\frac{e(H_i)}{v(H_i)}v(F_i)-e(F_i)\right)\\ &= \min_i\min_{F_i}\left(\frac{e(H)}{v(H)}v(F_i)-e(F_i)\right)\\ &\ge \kappa(H),\end{aligned}$$

since the F_i are also subgraphs of H. Thus, considering $\tau = 0$ and $\tau > 0$ together, we have

$$\text{Expression}(18) \leq C(\mathbb{E}W)\Phi n^{-\kappa(H)/d(H)}\mathbf{1}(\tau \neq 0). \tag{19}$$

Next we consider (17). If $\tau = 0$, then (17) is an empty sum, because γ is a path of length at most $\tau = 0$ and so $V(\delta_0)$ must intersect $V(\beta)$. Assuming $\tau > 0$, we bound the expectation of (17) by

$$\sum_{\substack{\beta \in \Gamma_\alpha^w \\ V(\beta)\cap V(\alpha)=\emptyset}} \sum_{i=0}^{k-1} \sum_{\substack{\delta_i \in \Gamma_{\alpha,i} \\ V(\delta_i)\cap V(\beta)=\emptyset}} \sum_{\gamma \in \Gamma(\delta_i,\beta)} \mathbf{P}\left(\delta_i \cup \gamma \cup \beta \subset G_\alpha(n, p(2-p))\right). \tag{20}$$

The choice of β gives $O(\mathbb{E}W)$. The choice of δ_i contributes

$$\max_i n^{v(\delta_i)-v(F_{\delta_i})} p^{e(\delta_i)-e(F_{\delta_i})} = O(\Psi),$$

where the F_{δ_i} are subgraphs of δ_i containing core(α). Note that one can map each vertex $v \in V(\gamma) \setminus V(\alpha \cup \delta_i \cup \beta)$ to the first edge of γ that contains v. Moreover, the first edge of γ containing a vertex in $V(\beta)$ must lie in $E(\gamma) \setminus E(\alpha \cup \delta_i \cup \beta)$ and is not the first edge that contains its other vertex. Therefore, the number of edges in $E(\gamma) \setminus E(\alpha \cup \delta_i \cup \beta)$ is greater than the number of vertices in $V(\gamma) \setminus V(\alpha \cup \delta_i \cup \beta)$ and the sum over γ contributes

$$\sup_{0 \leq i \leq \tau-1} n^i p^{i+1} = \frac{(np)^\tau}{n} \vee p.$$

We have shown that (20) is bounded by

$$C(\mathbb{E}W)\Psi\left(\frac{(np)^\tau}{n} \vee p\right)\mathbf{1}(\tau \neq 0). \tag{21}$$

Adding (11), (15), (19), and (21) shows that

$$\begin{aligned} & \mathbb{E}W_\alpha^+ + \mathbb{E}W_\alpha^- \\ \leq\ & C\Big\{ (\mathbb{E}W \vee 1)\, n^{-\kappa(H)/d(H)} + \mathbb{E}W\, \Phi\, n^{-\kappa(H)/d(H)}\mathbf{1}(\tau \neq 0) \\ & + \mathbb{E}W\, \Psi\, \frac{(np \vee 1)^\tau}{n} + \mathbb{E}W\, \Psi \left(\frac{(np)^\tau}{n} \vee p\right)\mathbf{1}(\tau \neq 0)\Big\}. \end{aligned}$$

Multiplying that bound by $\mathbb{E}W$ and adding the result to the sum of (8) and (9) gives the stated result. ■

Example: For the whisker graph in Figure 2, $\kappa(H) = (H) = 1$, $\Psi = O(1)$ and we have $\varepsilon_1 \leq Cn^7p^8$.

We now derive concrete compound Poisson approximations of W for local graphs at threshold. Let

$$\lambda := \sum_{i=1}^{\infty} \lambda_i.$$

A general bound on Q_1 was proved in [2]:

$$Q_1 \leq Ce^{\lambda}\left(\lambda^{-1} \wedge 1\right). \tag{22}$$

This bound is obviously not very useful as $\lambda \to \infty$. At threshold, however, $\mathbb{E}W$ is bounded as a function of n and $\lambda \leq \sum_i i\lambda_i = \mathbb{E}W < \infty$.

Theorem 1 *At the threshold $p = cn^{-v(H)/e(H)}$, there is a constant $C = C(c)$ such that*

$$\begin{aligned} d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{CP}(\boldsymbol{\lambda})) \;\leq\; & C\Big\{p^{e(H)} + n^{-v(H_0)} + n^{-\kappa(H)/d(H)} \\ & + n^{-\kappa(H)/d(H)}\mathbf{1}(\tau \neq 0) + \frac{(np \vee 1)^{\tau}}{n} \\ & + \left(\frac{(np)^{\tau}}{n} \vee p\right)\mathbf{1}(\tau \neq 0)\Big\}. \end{aligned}$$

This work opens at least two areas for further investigation:

- Are the Kolmolgorov-Smirnov distance bounds in [4] applicable to subgraph counts ? If so they would give compound Poisson approximations for local graphs for p larger than threshold.
- Subgraph counts may be expressed as incomplete U-statistics. An appropriate question for U-statistics would be "what conditions on sequences of incomplete U-statistics suffice to insure a compound Poisson limiting distribution"? Such conditions might be discovered using the insights about the compound Poisson approximation of subgraph counts contained in this paper.

References

[1] A. D. Barbour, Poisson convergence and random graphs, *Math. Proc. Camb. Math. Soc.*, **92**, 349 – 359.

[2] A. D. Barbour, L. H. Y. Chen, and W.-L. Loh, Compound Poisson approximation for nonnegative random variables via Stein's method, *Ann. Prob.*, **20**, 1843–1866.

[3] A. D. Barbour, L. Holst and S. Janson, *Poisson Approximation*, Oxford Science Publications, 1992.

[4] A. D. Barbour, S. Utev, Solving the Stein equation in compound Poisson approximation, to appear in *Adv. Appl. Prob.*

[5] B. Bollobás, *Random Graphs*, Academic Press, 1985.

[6] B. Bollobás, Threshold functions for small subgraphs, *Math. Proc. Camb. Phil. Soc.*, **90**, 197 – 206.

[7] B. Bollobás and J. Wierman, Subgraph counts and containment probabilities of balanced and unbalanced subgraphs in a large random graph, *Ann. New York Acad. Sci.*, VOLUME, 42–70.

[8] M. Roos, Stein-Chen method for compound Poisson approximation: the coupling approach, in *Prob. Th. and Math. Stat., Proc. 6th Vilnius Conference*, 645–660.

[9] P. Eichelsbacher and M. Roos, Compound Poisson approximation for dissociated random variables via Stein's method. Submitted.

[10] A. Ruciński, Small subgraphs of random graphs - a survey, in *Random Graphs '87*, Wiley.

[11] A. Ruciński and A. Vince, Balanced graphs and the problem of subgraphs of random graphs, *Congr. Num.*, **49**, 181 – 190.