A Poisson Approximation for Coloured Graphs Under Exchangeability

Annalisa Cerquetti and Sandra Fortini
Università Bocconi, Milano, Italy

Abstract

We introduce random graphs with exchangeable hidden colours and prove an asymptotic result on the number of times a fixed graph appears as a subgraph of such a random graph. In particular, we give necessary and sufficient conditions for the number of subgraphs isomorphic to a given graph to converge, under a negligibility assumption on the frequencies of colours. Moreover, we prove that the limiting law, when it exists, is a mixture of Poisson distribution. Our proofs rely on Stein-Chen method for Poisson approximations of sums of weakly dependent random variables. Finally, we discuss an application of the asymptotic result in Bayesian modeling.

AMS (2000) *subject classification.* Primary 60F05; secondary 05C80, 62F99.
Keywords and phrases. Exchangeability, Poisson approximation, random graphs, subgraph enumeration.

1 Introduction

The systematic study of random graphs was started by Erdős and Rényi in a series of seminal papers in 1950s and 1960s. In its simplest version, the classical (finite, undirected, loopless) *random graph $G(n,p)$*, is defined by taking a finite set of vertices $\{1, \ldots, n\}$ and then randomly selecting each of the $\binom{n}{2}$ possible edges with probability p independently of all other edges. As pointed out in Newman (2003), Erdős-Rényi model fails to describe many real-world networks, due to its intrinsic lack of edge correlation. In many different fields like molecular biology, ecology or information-technology, observed network structures frequently show *clustering*, i.e. vertices are more likely to be connected when they have a common neighbour. Hence in the last decade, an increased interest in random graph theory, has produced generalizations of the Erdős-Rényi model allowing correlation between edges (see Penman, 1998, Biggins, 2002, Biggins and Penman, 2003, Cannings

and Penman, 2003 and Söderberg, 2003). One of the most appealing is obtained by randomly *colouring* vertices according to a colour distribution and realizing edges independently with colour dependent probabilities.

An interesting class of results in random graph theory concerns the problem of counting the number of times a fixed subgraph H appears in a random graph. (For a general introduction to random graphs see e.g., Bollobas, 1985). Let $G(n,p)$ be the Erdős-Rényi model, and let W be the number of copies of H in $G(n,p)$. Barbour et al. (1992) establish an upper bound for the total variation distance between the law of W and a Poisson law with a suitable mean. Their proofs rely upon the *Stein-Chen method*, which is a general method to establish Poisson approximations for the sum of weakly dependent indicator random variables, with small occurrence probabilities. The main virtues of the Stein-Chen method is that it gives explicit upper bounds on the total variation distance.

In randomly coloured random graphs, subgraph enumeration was originally introduced by Janson (1986, 1987) as an alternative formulation of the famous *birthday problem* (Von Mises, 1939). Let Γ be a randomly coloured random graph obtained by colouring the vertices of a complete graph G, at random and independently, and then deleting the edges with different colours at the endpoints. If vertices represent people and colours represent people's birthdays, then an edge arises in Γ for every pair of people sharing the same birthday, so that the number of edges W corresponds to the number of pairs of people born in the same day. In Barbour et al. (1992), Theorem 5.G, the Stein-Chen method is used to approximate W by a Poisson random variable with mean $\lambda = \binom{n}{2} \sum_r p_r^2$, where p_r is the probability of being born on day r, and n is the number of people.

Here we extend the above result in two different directions. First we relax the hypothesis that vertices colours are independent by assuming that they are exchangeable. This extension is motivated by potential applications in Bayesian statistics (see Section 4). Actually, in order to exploit de Finetti's representation theorem, we make the stronger assumption that the sequence of colours can be extended to an infinite exchangeable sequence. Moreover we consider the number W of occurrences of a general connected subgraph H. In this framework we prove that, under suitable negligibility assumptions, the probability distribution of W can be well approximated by a mixture of Poisson laws, if the order of G is large.

The above mentioned results can be applied in a Bayesian perspective to model the number of coincidences arising in large networks. Consider a large

network in which a random characteristic is associated to each node, in such a way that different nodes are homogeneous with respect to it. It may be, for example, a particular choice or interest in a social network. Suppose that we want to fix a model for the number of coincidences among connected nodes in the network. In a Bayesian approach, if the distribution of the characteristic is not known, a prior distribution is assigned to it. According to de Finetti's representation theorem, this is equivalent to assume that the nodes are exchangeable with respect to the above mentioned characteristic. Our asymptotic results on random graphs with exchangeable colours imply that, if coincidences can be considered as exceptional events, a Poisson model with a random mean can be employed.

The paper is organized as follows. In Section 2 we fix some notations and introduce basic definitions. Moreover we state a property of random graphs with independent hidden colours exploited in the subsequent section. In Section 3 we present the asymptotic result with some examples, remarks and comments. In Section 4 we apply the asymptotic result to discuss a Bayesian version of the birthday problem.

2   Notations and Preliminary Results

Before introducing the definition of random graph with exchangeable hidden colours, we fix some notations. For basic definition on graph theory see e.g. Diestel (2000). Let $G = (V, E)$ be a graph: we can always take $V = \{1, \ldots, n\}$. Let $|G|$ denote the number of vertices of G. Subgraphs of a graph G will often be denoted by Greek letters α, β, etc. We write $\alpha \subset G$ to denote α being a subgraph of G. If a subgraph α of G is isomorphic to a fixed graph H, α is called a copy of H in G. The set of copies of H (in a fixed graph G) is denoted by $\mathcal{A}_H$. We use $\mathcal{B}_{\alpha,r}$ to denote the number of copies of H in G having r vertices in common with α, a particular copy of H:

$$\mathcal{B}_{\alpha,r} = \{\beta \in \mathcal{A}_H : |\beta \cap \alpha| = r\}. \tag{2.1}$$

Let

$$m_r = \frac{1}{|\mathcal{A}_H|} \sum_{\alpha \in \mathcal{A}_\mathcal{H}} |\mathcal{B}_{\alpha,r}|. \tag{2.2}$$

Suppose G is a graph with n vertices, and each vertex i receives a random colour X_i from a list of $k < \infty$ colours, where $(X_1, \ldots, X_n)$ have some joint distribution. A random subgraph Γ of G with hidden colours is the graph with vertex set $V(G)$ and edges those edges ij of G for which $X_i = X_j$.

More generally, if (G_n) is a sequence of graphs, with $|G_n| = n$, suppose each vertex of G_n is coloured with one of $k_n < \infty$ colours, the colour of vertex i being $X_{i,n}$ in G_n. We form a sequence (Γ_n) of random subgraphs with hidden colours in the same way.

The key case for us is when the $(X_{i,n})$, $1 \le i \le n$, are i.i.d. conditional on a random probability Q_n:

$$P(X_{1,n} \in A_1, \dots, X_{n,n} \in A_n | Q_n) = \prod_{i=1}^{n} Q_n(A_i). \tag{2.3}$$

Then de Finetti's theorem (see e.g., Aldous, 1985) implies that the joint distribution of the sequence of vertex colours is exchangeable:

$$P(X_{1,n} \in A_1, \dots X_{n,n} \in A_n) = P(X_{1,n} \in A_{\pi(1)}, \dots X_{n,n} \in A_{\pi(n)})$$

for all subsets A_s of the set of k colours for G_n, all π a permutation of $\{1, \dots, n\}$. Moreover, the finite sequence can be extended to an infinite exchangeable sequence. Note Q_n is defined by a k_n-vector $(Q_n(1), \dots, Q_n(k_n))$ which we write as $Q_n = (Q_{1,n}, \dots, Q_{k_n,n})$. In this case we call the resulting random graphs (Γ_n) random graphs with exchangeable hidden colours. A special case is when the colours are independent: these are a particular case of the random randomly coloured graphs introduced in Penman (1998) and Söderberg (2003) independently.

Let Γ be a random graph on an underlying graph G and let H be a graph. The number of copies of H in Γ is the random variable

$$W = \sum_{\alpha \in \mathcal{A}_H} I_\alpha,$$

where

$$I_\alpha = 1\{\alpha \subset \Gamma\}.$$

Here, for every B, $1(B)$ is the indicator function of B.

If Γ is a random graph with hidden colours and (X_i) is its sequence of random colours, then

$$W = \sum_{\alpha \in \mathcal{A}_H} 1(\cap_{ij \in \mathcal{V}(\alpha)} \{X_i = X_j\}).$$

Let the Poisson distribution with parameter $\lambda > 0$ be denoted by Po_λ. We include in the last class the degenerate distribution on 0 as the Po_0 distribution.

The asymptotic results of next section are based on the following theorem.

THEOREM 1. Let Γ be a random graph on G with independent and identically distributed colours $X_1, X_2, \ldots$ and colour space $\{1, \ldots, k\}$. Let H be a connected graph with $|H| > 1$. Then the number of copies of H in Γ satisfies

$$d_{TV}(\mathcal{L}(W), Po_\lambda) \leq 1 \wedge (1 - e^{-\lambda}) \left(\frac{\sum_{r=1}^{|H|-1} m_r + 1}{\mathcal{A}_H} \lambda + \sum_{r=1}^{|H|-1} m_r \frac{\sum_{i=1}^k q_i^{2|H|-r}}{\sum_{i=1}^k q_i^{|H|}} \right),$$

where d_{TV} denotes the total variation distance, $q_i = P(X_1 = i)$ for every i, and $\lambda = |\mathcal{A}_H| \sum_{i=1}^k q_i^{|H|}$.

PROOF. The random variables $\{I_\alpha : \alpha \in \mathcal{A}\}$ are dissociated (see Barbour et al. (1992) p. 34). Theorem 2.N in Barbour et al. (1992) then implies

$$d_{TV}(\mathcal{L}(W), Po_\lambda) \leq \frac{1 - e^{-\lambda}}{\lambda} \sum_{\alpha \in \mathcal{A}_H} \left((EI_\alpha)^2 + \sum_{r=1}^{|H|-1} \sum_{\beta \in \mathcal{B}_{\alpha,r}} ((EI_\alpha)^2 + (EI_\alpha I_\beta)) \right).$$

Moreover $E(I_\alpha) = \sum_{i=1}^k q_i^{|H|} = \lambda/|\mathcal{A}_H|$ while $E(I_\alpha I_\beta) = \sum_{i=1}^k q_i^{2|H|-r}$ if α and β have r common vertices $(1 \leq r \leq |H| - 1)$ $\qquad \square$

Let X_i be a sequence of i.i.d. random variables such that $P(X_i = j) = q_j$ for all i and j.

COROLLARY 2. Let $w : \{1, \ldots, k\}^n \to \{0, 1, \ldots\}$ be defined by

$$w(x_1, \ldots, x_n) = \sum_{\alpha \in \mathcal{A}_H} \prod_{ij \in \mathcal{V}(\alpha)} 1\{x_i = x_j\}.$$

and $\lambda = |\mathcal{A}_H| \sum_{i=1}^k q_i^{|H|}$. Then

$$|\sum_{x_1, \ldots, x_n \in \{1, \ldots, k\}} 1\{w(x_1, \ldots, x_n) \in A\} \prod_{i=1}^n q_{x_i} - Po_\lambda(A)|$$

$$\leq 1 \wedge (1 - e^{-\lambda}) \left(\frac{\sum_{r=1}^{|H|-1} m_r + 1}{\mathcal{A}_H} \lambda + \sum_{r=1}^{|H|-1} m_r \frac{\sum_{i=1}^k q_i^{2|H|-r}}{\sum_{i=1}^k q_i^{|H|}} \right),$$

(2.4)

3   Poisson Approximation

In this section we use Corollary 2 to show that, under a suitable negligibility assumption, the asymptotic distribution of (W_n) is a mixture of Poisson laws.

Let μ be a probability distribution on the Borel sigma-algebra on $[0, +\infty)$. A probability measure ν satisfying

$$\nu(A) = \int_{[0,+\infty)} Po_\lambda(A) d\mu(\lambda)$$

is called a mixture of Poisson laws with mixing measure μ. The following theorem asserts that ν uniquely determines μ.

THEOREM 3. Let ν' and ν'' be mixtures of Poisson laws with mixing measures μ' and μ'', respectively. If $\nu'' = \nu'$, then $\mu'' = \mu'$.

The proof is obvious (see e.g., Kallenberg, 1986, Corollary 3.2).

In the following, we will assume that (Q_n) satisfies the following negligibility assumption:

$$(m_{r,n} + 1)\frac{\sum_{i=1}^{k_n} Q_{i,n}^{2|H|-r}}{\sum_{i=1}^{k_n} Q_{i,n}^{|H|}} \xrightarrow{P} 0 \text{ as } n \to \infty, \text{ for every } r = 1, \ldots, |H| - 1,$$

(3.1)

where $\xrightarrow{P}$ means convergence in probability,

$$m_{r,n} = \frac{1}{|\mathcal{A}_{H,n}|} \sum_{\alpha \in \mathcal{A}_{H,n}} |\mathcal{B}_{\alpha,r,n}|$$

(3.2)

and

$$\mathcal{B}_{\alpha,r,n} = \{\beta \in \mathcal{A}_{H,n} : |\beta \cap \alpha| = r\}.$$

REMARK 1. A sufficient condition for (3.1) is

$$(m_{r,n} + 1)^{1/(|H|-r)} \max_i Q_{i,n} \xrightarrow{P} 0 \text{ for every } r = 1, \ldots, |H| - 1,$$

(3.3)

as $n \to \infty$.

REMARK 2. Assumption (3.1) implies that $(I_{\alpha,n})$ are uniformly asymptotically negligible, that is $\max_\alpha P(I_{\alpha.n} = 1)$ goes to zero, as $n \to \infty$. In fact, since $\sum_{i=1}^{k_n} Q_{i,n} = 1$,

$$\sum_{i=1}^{k_n} Q_{i,n}^{2|H|-r} \geq \left(\sum_{i=1}^{k_n} Q_{i,n}^{|H|}\right)^{\frac{2|H|-r-1}{|H|-1}} \geq \left(\sum_{i=1}^{k_n} Q_{i,n}^{|H|}\right)^2.$$

Moreover, if $\overline{\alpha}$ is a fixed copy of H in G_n, $E(I_{\overline{\alpha},n}) = \sum_{i=1}^{k_n} Q_{i,n}^{|H|}$. Hence,

$$\max_{\alpha \in \mathcal{A}_{H,n}} P(I_{\alpha,n} = 1) = E(P(I_{\overline{\alpha},n} = 1|Q_n))$$

$$= E(\sum_i Q_{i,n}^{|H|}) \leq E\left(\frac{\sum_{i=1}^{k_n} Q_{i,n}^{2|H|-r}}{\sum_{i=1}^{k_n} Q_i^{|H|}}\right)$$

If (3.1) holds, the last term in the inequality tends to zero, as $n \to \infty$. Notice that the number of colours k_n must go to infinity for (3.1) to hold.

On the other hand, the probability distribution of (W_n) might fail to converge to a mixture of Poisson laws, if (3.1) does not hold, even under the hypothesis that $(I_{\alpha,n})$ are uniformly asymptotically negligible. See Example1.

THEOREM 4. Let (Γ_n) be a sequence of random graphs with exchangeable hidden colours on a sequence of graphs (G_n), with $|G_n| = n$. Let H be a connected graph with $|H| > 1$ and let $\mathcal{A}_{H,n}$, $m_{r,n}$ and Q_n be defined as above with $|\mathcal{A}_{H,n}| \to \infty$ as $n \to \infty$. Suppose that (3.1) holds and let

$$\Lambda_n = |\mathcal{A}_{H,n}| \sum_i Q_{i,n}^{|H|}. \tag{3.4}$$

Then (W_n) converges in distribution if and only if (Λ_n) converges in distribution. In this case, the limiting distribution of (W_n) is a mixture of Poisson laws, the mixing measure being the limiting law of (Λ_n). Moreover the representation is unique.

PROOF. Let

$$w_n(x_1, \ldots, x_n) = \sum_{\alpha \in \mathcal{A}_{H,n}} \prod_{i,j \in \alpha} 1\{x_i = x_j\}.$$

Then $W_n = w_n(X_{1,n}, \ldots, X_{n,n})$. Moreover

$$P(W_n \in A|Q_n) = \sum_{x_1, \ldots, x_n} 1\{w_n(x_1, \ldots, x_n) \in A\} \prod_{i=1}^n Q_{x_i,n} \quad \text{P-a.s.}$$

It follows from (2.3) and (2.4) that, for every Borel set A,

$$|P(W_n \in A|Q_n) - Po_{\Lambda_n}(A)|$$

$$\leq 1 \wedge \left(\left(\sum_{r=1}^{|H|-1} m_{r,n} + 1\right) \sum_{i=1}^{k_n} Q_{i,n}^{|H|} + \sum_{r=1}^{|H|-1} m_{r,n} \frac{\sum_{i=1}^{k_n} Q_{i,n}^{2|H|-r}}{\sum_{i=1}^{k_n} Q_{i,n}^{|H|}}\right)$$

$$\leq 1 \wedge \left(\sum_{r=1}^{|H|-1}(m_{r,n} + 1) \sum_{i=1}^{k_n} Q_{i,n}^{|H|} + \sum_{r=1}^{|H|-1} m_{r,n} \frac{\sum_{i=1}^{k_n} Q_{i,n}^{2|H|-r}}{\sum_{i=1}^{k_n} Q_{i,n}^{|H|}}\right) \quad \text{P-a.s.}$$

$$\tag{3.5}$$

By (3.1) and Remark 2, we see that the random vector

$$\left((m_{1,n}+1)\sum_{i=1}^{k_n}Q_{i,n}^{|H|},\ldots,(m_{|H|-1,n}+1)\sum_{i=1}^{k_n}Q_{i,n}^{|H|},\right.$$

$$\left.m_{1,n}\frac{\sum_{i=1}^{k_n}Q_{i,n}^{2|H|-1}}{\sum_{i=1}^{k_n}Q_{i,n}^{|H|}},\ldots,m_{|H|-1,n}\frac{\sum_{i=1}^{k_n}Q_{i,n}^{|H|+1}}{\sum_{i=1}^{k_n}Q_{i,n}^{|H|}}\right)$$

converges in distribution to the zero vector, as $n\to\infty$. Since

$$1\wedge\left(\sum_{r=1}^{|H|-1}(m_{r,n}+1)\sum_{i=1}^{k_n}Q_{i,n}^{|H|}+\sum_{r=1}^{|H|-1}m_{r,n}\frac{\sum_{i=1}^{k_n}Q_{i,n}^{2|H|-r}}{\sum_{i=1}^{k_n}Q_{i,n}^{|H|}}\right)$$

is a bounded, continuous function of it,

$$|P(W_n\in A)-E(Po_{\Lambda_n}(A))|\le E|P(W_n\in A|Q_n)-Po_{\Lambda_n}(A)|\to0\quad(3.6)$$

as $n\to\infty$.

If Λ_n converges in distribution to a random variable Λ, then

$$|E(Po_{\Lambda_n}(A))-E(Po_\Lambda(A))|\to0$$

and, therefore the probability distribution of W_n converges, as $n\to\infty$, to $E(Po_\Lambda)$.

Conversely, suppose that

$$W_n\overset{d}{\to}W,$$

as $n\to\infty$. For every n, let μ_n denote the probability distribution of Λ_n. Let us prove that (μ_n) is tight. By Chebyshev's inequality, for every $\lambda>0$,

$$Po_\lambda[0,\lambda/2]\le\frac{4}{\lambda}.$$

Suppose that (μ_n) is not tight and let $a<1$ be such that, for every $R>4$, there exists an increasing sequence of integers (n_m), depending on R, such that, for every m, $p_{n_m}=P(\Lambda_{n_m}\le2R)<a$. Then

$$Po_{\Lambda_{n_m}}[0,R]\le Po_{\Lambda_{n_m}}[0,R]1\{\Lambda_{n_m}<2R\}+Po_{\Lambda_{n_m}}[0,R]1\{\Lambda_{n_m}\ge2R\}$$
$$\le1\{\Lambda_{n_m}<2R\}+Po_{\Lambda_{n_m}}[0,\Lambda_{n_m}/2]1\{\Lambda_{n_m}\ge2R\}$$
$$\le1\{\Lambda_{n_m}<2R\}+4/\Lambda_{n_m}1\{\Lambda_{n_m}\ge2R\})$$
$$\le1\{\Lambda_{n_m}<2R\}+2/R1\{\Lambda_{n_m}\ge2R\}$$

Since $R > 4$,

$$E(Po_{\Lambda_{n_m}}[0, R]) \leq p_{n_m} + 2/R(1 - p_{n_m}) \leq (a+1)/2.$$

From (3.6),
$$P(W_{n_m} \leq R) - E(Po_{\Lambda_{n_m}}[0, R]) \to 0,$$

as $m \to \infty$. Let m_0 be such that, for every $m \geq m_0$,

$$|P(W_{n_m} \leq R) - E(Po_{\Lambda_{n_m}}[0, R])| \leq 1/4 - a/4.$$

Then, for every $m \geq m_0$,

$$P(W_{n_m} \leq R) \leq (3+a)/4 < 1.$$

Concluding, if μ_n is not tight, there exists $b < 1$ such that, for every $R > 4$, $P(W_n \leq R) \leq b$ for some n. Hence W_n fails to converge in distribution, which contradicts the hypothesis.

We have thus proved that (μ_n) is tight. Recall (Billingsley, 1995, Theorem 25.10) that this implies every subsequence has a subsequence which converges to a probability measure, and the corollary that, if (μ_n) is tight, with each subsequence that converges weakly at all converging weakly to the probability measure μ, then (μ_n) converges in distribution to μ. Let $(\mu_{n'})$ and $\mu_{n''}$ be subsequences converging to μ' and μ'', respectively. Then, by the first part of the proof, the limiting distributions of $W_{n'}$ and $W_{n''}$ are $\int Po_\lambda \mu'(d\lambda)$ and $\int Po_\lambda \mu''(d\lambda)$, respectively. Since W_n converges in distribution,

$$\int Po_\lambda \mu'(d\lambda) = \int Po_\lambda \mu''(d\lambda).$$

It follows from Theorem 1 that $\mu' = \mu''$.   □

Assumption (3.1) of Theorem 4 can not be suppressed, as proved in Example 1.

EXAMPLE 1. For every n, let K_n be the complete graph with n vertices and $k_n = n^3$. For every n, let Γ_n be a random graph on G_n with independent and identically distributed colours $X_{1,n}, \ldots, X_{n,n}$, taking values $1, \ldots, n^3$ with probabilities, $1/n, 1/(n^3+n^2+n), 1/(n^3+n^2+n), \ldots, 1/(n^3+n^2+n)$, respectively. Let H be an edge. Then $|\mathcal{A}_{H,n}| = \binom{n}{2}$ and $\Lambda_n = \binom{n}{2} \sum_{j=1}^{n^3} Q_{j,n}^2 \to 1$, as $n \to \infty$. On the other hand, (3.1) does not hold in this case, since

$$(m_{1,n} + 1)\frac{\sum_{i=1}^{n^3} Q_{i,n}^3}{\sum_{i=1}^{n^3} Q_{i,n}^2} = (2n-3)\frac{n^{-3} + (n^3-1)(n^3+n^2+n)^{-3}}{n^{-2} + (n^3-1)(n^3+n^2+n)^{-2}} \not\to 0,$$

as $n \to \infty$. We show the limiting distribution, if it exists, is not a mixture of Poisson laws by showing that, as $n \to \infty$,

$$P(W_n = 1) \to e^{-1}/2, \ P(W_n = 2) \to 0. \tag{3.7}$$

To prove (3.7), let C_n be the number of vertices with colour 1 in Γ_n. Then $C_n \sim Bin(n, 1/n)$ so asymptotically $C_n \sim Po_1$. Now, for $w \leq 2$

$$P(W_n = w) = \sum_{c=0}^{2} P(W_n = w | C_n = c) P(C_n = c)$$

Let $W_n^{(c)}$ be the number of edges in a random graph with independent colours, on a complete graph with $n - c$ vertices, with all $n^3 - 1$ colours being equally likely. It is easy to check directly (or by Theorem 3) that $W_n^{(c)}$ has a limiting Po_0 distribution so

$$P(W_n = w | C_n = c) = P(W_n^{(c)} = w) \to 0 \quad \text{for } w = 1, 2, c = 0, 1$$
$$P(W_n = 1 | C_n = 2) = P(W_n^{(c)} = 0) \to 1$$
$$P(W_n = 2 | C_n = 2) = P(W_n^{(c)} = 1) \to 0,$$

as $n \to \infty$.

EXAMPLE 2. Let G_n be the complete graph with n vertices and let Γ_n be a random graph with exchangeable hidden colours taking values $1, \ldots, n^2$. Let (a_n) be a sequence of real numbers such that

$$\left. \begin{array}{l} \sum_{i=1}^{n^2} \frac{a_i}{n^2} \to \lambda_1 > 0 \\[2mm] \sum_{i=1}^{n^2} \frac{a_i^2}{n^2} \to \lambda_2 > 0 \end{array} \right\} \quad \text{as } n \to \infty. \tag{3.8}$$

Suppose that the random vector $(Q_{1,n}, \ldots, Q_{n^2,n})$ has Dirichlet distribution with parameters $(a_1, \ldots, a_{n^2})$ (see e.g. Wilks, 1962). We prove that the distribution of number of edges in Γ_n, W_n, converges to the $Po_{(\lambda_1+\lambda_2)/2\lambda_1^2}$. This will follow from Theorem 4, if we can prove that

$$(m_{1,n} + 1)\frac{\sum_{i=1}^{n^2} Q_{i,n}^3}{\sum_{i=1}^{n^2} Q_{i,n}^2} \xrightarrow{P} 0 \tag{3.9}$$

and

$$|\mathcal{A}_{H,n}| \sum_{i=1}^{n^2} Q_{i,n}^2 \xrightarrow{P} \frac{\lambda_1 + \lambda_2}{2\lambda_1^2} \tag{3.10}$$

as $n \to \infty$. (3.9) and (3.10) will follow from the fact that there is a sequence $(Y_{i,n})_{i=1,\ldots,n^2,n=1,2,\ldots}$ of i.i.d. random variables with Gamma distribution such that $Q_{i,n} = Y_{i,n}/\sum_{i=1}^{n^2} Y_{i,n}$ (see Wilks, 1962, Theorem 7.7.1). Hence

$$(m_{1,n}+1)\frac{\sum_{i=1}^{n^2} Q_{i,n}^3}{\sum_{i=1}^{n^2} Q_{i,n}^2} = \frac{2(n-2)+1}{n^2} \cdot \frac{\frac{\sum_{i=1}^{n^2} Y_{i,n}^3}{n^2}}{\frac{\sum_{i=1}^{n^2} Y_{i,n}}{n^2} \cdot \frac{\sum_{i=1}^{n^2} Y_{i,n}^2}{n^2}}$$

and

$$|\mathcal{A}_{H,n}| \sum_{i=1}^{n^2} Q_{i,n}^2 = \frac{n-1}{2n} \cdot \frac{\frac{\sum_{i=1}^{n^2} Y_{i,n}^2}{n^2}}{\left(\frac{\sum_{i=1}^{n^2} Y_{i,n}}{n^2}\right)^2}.$$

Since

$$E\left(\frac{\sum_{i=1}^{n^2} Y_{i,n}^2}{n^2}\right) = \frac{\sum_{i=1}^{n^2} a_i(a_i+1)}{n^2} \to \lambda_1 + \lambda_2$$

and

$$V\left(\frac{\sum_{i=1}^{n^2} Y_{i,n}^2}{n^2}\right) = \frac{\sum_{i=1}^{n^2} a_i(a_i+1)(4a_i+6)}{n^4} \to 0,$$

as $n \to \infty$, then $\frac{\sum_{i=1}^{n^2} Y_{i,n}^2}{n^2}$ converges in probability to $\lambda_1 + \lambda_2$, as $n \to \infty$. Moreover

$$E\left(\frac{\sum_{i=1}^{n^2} Y_{i,n}}{n^2}\right) = \frac{\sum_{i=1}^{n^2} a_i}{n^2} \to \lambda_1$$

and

$$V\left(\frac{\sum_{i=1}^{n^2} Y_{i,n}}{n^2}\right) = \frac{\sum_{i=1}^{n^2} a_i}{n^4} \to 0,$$

as $n \to \infty$. Hence $(\frac{\sum_{i=1}^{n^2} Y_{i,n}}{n^2})^2 \xrightarrow{P} \lambda_1^2$, as $n \to \infty$. Since $\frac{\sum_{i=1}^{n^2} Y_{i,n}^3}{n^2}$ is unlikely to exceed the mean of $Y_{i,n}^3$ by much, (3.9) follows as the $2(n-2)/n^2$ term pulls the expression to zero.

In the above example, the random parameter of the limiting Poisson law has a degenerate probability distribution. Hence the mixture distribution reduces to a Poisson law. The following example shows a situation in which the limiting law of W_n is a proper mixture of Poisson laws.

EXAMPLE 3. Let s be a fixed positive integer and let K_n be the complete graph with n vertices. Consider a sequence of random graphs (Γ_n) on K_n with colours $\{1, \ldots, sc_n\}$, where $c_n = \lfloor n^2/s \rfloor$. Let $(Z_1, \ldots, Z_s)$ be a random vector with $Z_i \geq 0$ for every i and $\sum_{i=1}^{s} Z_i = 1$ and let

$$Q_{i,n} = Z_j/c_n \text{ for every } i = (j-1)c_n + 1, \ldots, jc_n \ (j = 1, \ldots, s).$$

Let W_n be the number of edges in Γ_n. Then $|H| = 2$, $|\mathcal{A}_{H,n}| = \binom{n}{2}$ and $m_{1,n} = 2(n-2)$. Moreover (3.1) holds. In fact

$$(m_{1,n} + 1)\frac{\sum_{i=1}^{sc_n} Q_{i,n}^3}{\sum_{i=1}^{sc_n} Q_{i,n}^2} = (2(n-2)+1)\frac{c_n \sum_{j=1}^{s} Z_j^3/c_n^3}{c_n \sum_{j=1}^{s} Z_j^2/c_n^2} \xrightarrow{P} 0$$

as $n \to \infty$. Since

$$\Lambda_n = \binom{n}{2}\sum_{i=1}^{sc_n} Q_{i,n}^2 = \binom{n}{2}\sum_{j=1}^{s} c_n \frac{Z_j^2}{c_n^2}$$

converges in distribution to $\Lambda = s\sum_{j=1}^{s} Z_j^2/2$, as $n \to \infty$, the limiting distribution of the number of edges in Γ_n converges to a mixture of Poisson laws, the mixing measure being the probability distribution of Λ.

4 A Bayesian Version of the Birthday Problem

As a variant of the birthday problem, consider a group of n people with various social ties, and let W be the number of pairs who are acquainted and share the same birthday. As suggested in Barbour et al. (1992), this problem can be tackled by random graphs with hidden colours. In a graph G with N edges, let vertices represent people and edges represent social ties. Consider the random graph obtained by colouring vertices with people's birthdays and then deleting edges with different colours at the endpoints. The number of remaining edges represents the number W of pairs of people who are acquainted and have the same birthdays. If the birthday distribution is uniform on k days, k and N are both large but $\sqrt{N}/k$ is small, then W has distribution approximately Poisson $(\frac{N}{k})$ (see Barbour et al., 1992, pp. 104-107). Hence $P(\text{no matches}) \doteq e^{-\lambda}$, where $\lambda = N/k$. If $k = 365$, the chance of a match is approximately 0.5 for $N = 253$. In the case of a complete graph, 253 edges correspond to 23 vertices, which is the classical answer to the birthday problem, i.e. 23 persons are required to obtain a probability 0.5 of a match. Diaconis and Holmes (2002) point out that, due to seasonal and other effects, the probability Q_i of a person being born on day i may differ

from $1/k$. Therefore, in a Bayesian approach, one should treat the Q_i's as unknown and specify a prior distribution for the random vector $(Q_1, \ldots, Q_k)$.

Diaconis and Holmes (2002) give an explicit formula for the probability of a match under a symmetric Dirichlet prior with parameters $(c, \ldots, c)$, when the graph G is complete. If c is small, then the prior distribution assigns high probability far from $(1/k, \ldots, 1/k)$, and this makes a match more likely. For example, under a uniform prior, $(c = 1)$, only 17 individuals are needed for a 50-50 chance of a match.

Suppose now that G has N edges but is otherwise arbitrary. If the assumptions of Theorem 4 hold, then the number of matches has an approximate Poisson distribution with random parameter

$$\Lambda = N \sum_i Q_i^2 = \frac{N}{k} + N \sum_i \left(Q_i - \frac{1}{k} \right)^2 \qquad (4.1)$$

It is easily seen from (4.1) that Λ has both a deterministic component $\lambda = N/k$, corresponding to the basic assumption of the classical birthday problem, and a random component. It follows that we can write

$$\Lambda = \lambda(1 + V)$$

where $V = k \sum_i (Q_i - 1/k)^2$ is a non-negative random variable whose distribution depends only on the prior and

$$Pr(\text{no matches}) = E(e^{-\Lambda}) = e^{-\lambda} E(e^{-\lambda V}).$$

For example if V has a Gamma $(a, b, 0)$ distribution (see e.g. Johnson and Kots, 1970)

$$Pr(\text{no matches}) = e^{-\lambda}(1 + \lambda b)^{-a}.$$

In what follows, we construct a prior distribution which satisfies the hypotheses of Theorem 4, and such that V follows, approximately, a Gamma $(a, b, 0)$ distribution. Let

$$Q_i = \frac{Y_i}{\sum_{j=1}^n Y_j} \ (i = 1, \ldots, k)$$

with $Y_1, \ldots, Y_k$ i.i.d. Generalized Gamma$(a/k, 2, 1/k, \sqrt{b/k})$ (see e.g. Johnson and Kots, 1970), let N and k diverge in such a way that $N/k \to \lambda$ and let $(m_1 + 1)/\sqrt{N} \to 0$ as $N \to \infty$ (see (2) for the definition of m_1). Then, for $k \to \infty$,

$$k \sum_i Q_i^2 \xrightarrow{d} 1 + V \qquad (4.2)$$

with $V \sim Gamma(a, b, 0)$ and

$$(m_1 + 1)\frac{\sum_i Q_i^3}{\sum_i Q_i^2} \xrightarrow{P} 0. \tag{4.3}$$

In fact,

$$k \sum_i Q_i^2 = k\frac{\sum_i Y_i^2}{(\sum_i Y_i)^2}.$$

Moreover, $\sum_i Y_i \xrightarrow{P} 1$ as $k \to \infty$, $k \sum_i (Y_i - 1/k)^2 \sim Gamma(a, b, 0)$ and

$$k \sum_i Y_i^2 = k \sum_i \left(Y_i - \frac{1}{k}\right)^2 + 2\sum_i Y_i - 1,$$

proving (4.2). Now, (4.3) holds since

$$(m_1 + 1) \sum_i Q_i^3 = (m_1 + 1)\frac{k \sum_i Y_i^3}{(\sum_i Y_i)(k \sum_i Y_i^2)}$$

and $E(k \sum_i Y_i^3) = o(1/(m_1 + 1))$, as $k \to \infty$.

The following table shows how varying a and b affects the N needed to obtain probability 0.5 of a match.

	$b = 0.2$	0.04	0.008	0.0016	0.00032
$a = 1$	213	243	251	253	253
2	183	234	249	252	253
5	129	211	243	251	253
10	86	181	234	249	252

Notice that the answer to the classical birthday problem is recovered for $b \to 0$.

Acknowledgements. We wish to thank Professor Cifarelli for his valuable comments.

References

ALDOUS, D.J. (1985). *Exchangeability and Related Topics.* École d'Été de Probabilitiés de Saint-Flour XIII - Lecture Notes in Mathematics **1117**. Springer-Verlag, Berlin.

Barbour, A.D., Holst, L. and Janson, S. (1992). *Poisson Approximations*. Oxford University Press, Oxford.

Biggins, J.D. (2003). *Large deviations for mixtures.* Tech. Rep. University of Sheffield, U.K.

Biggins, J.D. and Penman, D.B. (2003). Large deviations in randomly colored random graphs. Preprint, University of Sheffield, U.K.

Billingsley, P. (1995). *Probability and Measure.* Wiley, New York.

Bollobás, B. (1985). *Random Graphs.* Academic Press, New York.

Cannings, C. and Penman, D.B. (2003). Models of random graphs and their applications. In *Handbook of Statistics* **21**. *Stochastic Processes: Modelling and Simulation.* D.N. Shanbhag and C.R. Rao, eds. Elsevier, 51–91.

Chen, L.H.Y. (1975). Poisson approximations for dependent trials. *Ann. Probab.*, **3**, 534–545.

Diaconis, P. and Holmes, S. (2002) A Bayesian peek into Feller Volume I. *Sankhyā Ser. A*, **64**, 820–841.

Diestel, R. (2000). *Graph Theory*, Second ed. Springer-Verlag, New York.

Erdős, P. and Rényi, A. (1960). On the evolution of random graph. *Publ. Math. Inst. Hungar. Acad. Sci.* **5**, 17.

Janson, S. (1986). Birthday problems, randomly colored graphs and Poisson limits of sums of dissociated variables. Uppsala University, Department of Mathematics, Report No. 1986:16

Janson, S. (1987) Poisson convergence and Poisson processes with applications to random graphs. *Stoch. Proc. Appl.*, **26**, 1–30.

Johnson, N.L. and Kotz, S. (1970). *Continuous Univariate Distributions - I.* Wiley, New York.

Kallenberg, O. (1986). *Random Measures.* Academic Press, New York.

Newman, M.E.J. (2003). Random graphs as models of networks. In *Handbook of Graphs and Networks*, S. Bornholdt and H. G. Schuster, eds. Wiley-VCH, Berlin.

Penman, D.B. (1998). Random graphs with correlation structure. Ph.D Thesis, University of Sheffield.

Söderberg, B. (2003). Random graphs with hidden color. *Phys. Rev. E* **68** 015102.

Von Mises, R. (1939). Über aufteilungs und besetzungswahrscein lichkeiten, *Rev. Fac. Sci. Istambul*, **4**, 145–163.

Wilks, S.S. (1962). *Mathematical Statistics.* Wiley, New York.

Annalisa Cerquetti and Sandra Fortini
Istituto di Metodi Quantitativi
Universitá Bocconi
Viale Isonzo 25
20135 Milano, Italy
E-mail: Annalisa.Cerquetti@unibocconi.it
Sandra.Fortini@unibocconi.it