A Poisson approximation for colored graphs under exchangeability
Annalisa Cerquetti and Sandra Fortini
IQM - Bocconi University, Milan, Italy
annalisa.cerquetti, sandra.fortini@unibocconi.it

6th BS/IMSC Barcelona, July 26-31, 2004

Introduction
A large class of results in graphs theory concerns the problem of counting the number of times a certain substructure appears in a random graph; see e.g. Bollobás (1985).

Let \( G(n, p) \) be the Erdős random graph model, 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.

**Fig. 1:** Samples of Erdős random graphs.

For a fixed graph \( H \), consider the number \( W_n \) of copies of \( H \) in \( G(n, p) \). It is known that if \( p \) is such that the expected number of \( W_n \) converges to some finite value, then \( W_n \) has asymptotically a Poisson distribution; see Barbour et al. (1992) and Bollobás (1985).

Results of this kind rely upon the Chen-Stein method, which is a general method to establish Poisson approximations for sums of weakly dependent indicator random variables, with small occurrence probabilities; see Chen (1975).

As pointed out in Newman (2003), Erdős standard model generally fails to describe many real-world networks for its intrinsic lack of edge correlation. In fact, 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 neighbor.

One of the most appealing generalizations of the Erdős basic model, allowing correlation among edges, is obtained by randomly coloring vertices and realizing edges independently with color dependent probabilities; see e.g. Penman (1998).

Subgraphs enumeration in randomly colored random graphs was originally introduced in Janson (1986) as an alternative formulation of the famous birthday problem. If the \( n \) vertices represent people and vertex color \( X_i \) denotes birthday of person \( i \), then the number of pairs of people sharing the same birthday is the number \( W_n \) of edges in a randomly colored graph, where edges arise between vertices sharing the same color. In this framework Janson (1986) gives an upper bound to the total variation distance between the law of \( W_n \) and a Poisson distribution. The above approximation allows to prove a limit result for the counting problem in randomly colored random graphs when \( H \) consists of one edge connecting two vertices and the dimension of the graph \( n \) tends to infinity. In fact, if \( \lambda_n \) denotes the probability of color \( \lambda \),

\[ \mathbb{E}(W_n) = \lambda \text{ and } \frac{\sum \lambda_i^2}{\sum \lambda_i} = 0 \]

as \( n \to \infty \), then \( W_n \) has a limiting Poisson(\( \lambda \)) distribution.

**Random graphs with exchangeable colors**

Let \( G \) be a graph with \( n \) vertices and let \( K_n = \{a_1, a_2, \ldots, a_n\} \) be a finite set called color space. Let \( X_{a_1}, \ldots, X_{a_n} \) be an infinite sequence of random variables taking values in \( K_n \) and satisfying the exchangeability condition

\[ P(X_{a_1} = x_1, \ldots, X_{a_m} = x_m) = P(X_{a_1} = x_{a(1)}, \ldots, X_{a_m} = x_{a(m)}) \]

for every \( m \) and every permutation \( \pi \).

By de Finetti’s representation theorem

- \( X_{a_1}, X_{a_2}, \ldots \) are conditionally independent and identically distributed, given a random vector \( Q \) = \( (Q_{a_1}, Q_{a_2}, \ldots) \) and

\[ Q_a = P(X_a = i | Q) \quad i = 1, \ldots \]

A random graph with exchangeable colors \( G \) is obtained by

- assigning color \( X_{a_1} \) to the \( y \)-th vertex of \( G \)
- deleting edges with different colors at the endpoints.

**Fig. 2:** Samples of random graphs with exchangeable colors (\( G \), complete).

Exchangeability implies that all the configurations obtained, from a base one, by interchangeing vertices colors have the same probability.

Random graphs with exchangeable colors can be used to model clusters of networks nodes in Bayesian statistics. Let the partition be based on a characteristic vector \( \lambda \) being the complex network considered or on a characteristic vector \( \lambda \) being the complex network considered.

**Poisson approximation**

For every \( n \), let \( \lambda_n \) be a random graph with exchangeable colors on an underlying graph \( G_n \). Let \( H_n \) be a fixed connected graph with \( v \geq 2 \) vertices.

Suppose that the number of copies of \( H_n \) in \( G_n \) goes to infinity as \( n \) tends to infinity. Then, for every \( n \), let \( W_n \) be the number of copies of \( H_n \) in \( G_n \) and let \( \lambda_n = \mathbb{E}(W_n | Q) \).

The following limit result holds under the negligibility assumption (1) specified in next section.

**Theorem 1.** The sequence of random variables \( W_n \) converges in distribution if and only if \( \lambda_n \) converges in distribution. In this case, the limiting distribution of \( W_n \) is a mixture of Poisson laws: as \( n \to \infty \)

\[ \mathbb{P}(W_n) = \int_{\mathbb{R}_+} e^{\lambda_n} \frac{\lambda_n^{w-1}}{w!} d\lambda_n \]

the mixing measure \( \nu \) being the limiting law of \( \lambda_n \).

Moreover the representation is unique.

The proof of Theorem 1 relies on the Chen-Stein Poisson approximation theory.

**Remark 1.** Notice that \( W_n \) is the sum of indicators random variables

\[ W_n = \sum_{a} I_{a,n} \]

where the sum extends over the set of copies of \( H \) in \( G_n \) and \( I_{a,n} \) equals one if \( c_{a,n} \in C_n \). Hence the convergence of (a) exchangeable (and non-exchangeable) random variables and we expect that some negligibility assumption is necessary for the convergence of \( W_n \) to hold.

**Remark 2.** In Diaconis and Holmes (2002) asymptotic results for the birthday problem \( (H = 2) \) in a Bayesian framework, are obtained, without resorting to graphs theory and assuming a Dirichlet distribution for \( Q \). In this case the number of matches has a limiting Poisson distribution.

**The negligibility assumption**

For every set \( A \), let \( |A| \) denote the number of points in \( A \). Let

- \( A_{n,H} \) be the set copies of \( H \) in \( G_n \)
- \( B_{n,H} = \{ (i, j) \in A_{n,H} : \{i, j\} \cap H = \emptyset \} \) be the set of copies of \( H \) in \( G_n \) having \( r \) vertices in common with \( H \)
- \( m_a = \frac{1}{\lambda_n} \sum_{v \in \lambda_n} I_{a,v} \) be the mean number of copies of \( H \) in \( G_n \), having \( r \) vertices in common with \( H \).

The negligibility assumption states that

\[ (m_a + 1)^{-1} \sum_{v \in \lambda_n} I_{a,v} \to 0 \quad (1) \]

as \( n \to \infty \) for every \( r = \{ |H| - 1 \} \).

**Remark 3.** A sufficient condition for (1) is

\[ (m_a + 1)^{-1} \sum_{v \in \lambda_n} I_{a,v} \to 0 \quad (2) \]

as \( n \to \infty \). On the other hand, the probability distribution of \( W_n \) might fail to converge to a mixture of Poisson laws, if (1) does not hold, even under the hypothesis that \( \lambda_n \) are uniformly asymptotically negligible.

**Example**

Let \( s \) be a fixed positive integer and let \( \Gamma_n \) be a random graph on a complete graph with \( s \) vertices, colored with \( s^2 \) different colors \( 1, \ldots, s^2 \).

Let \( Z_{1,n}, \ldots, Z_{s^2,n} \) be positive random variables with \( \sum Z_{a,n} = 1 \) and let the random probabilities of colors \( Q_{a,n} \) be defined by the following rule:

- the random probability of colors \( 1, \ldots, s^2 \) is given by \( Z_{a,n}/s^2 \);
- the random probability of colors \( s^2 + 1, \ldots, 2s^2 \) is given by \( Z_{a,n}/s^2 \);
- ... the random probability of colors \( s(s-1)s^2+1, \ldots, s^2 \) is given by \( Z_{a,n}/s^2 \).

By Theorem 1 the limiting distribution of the number of edges in \( \Gamma_n \) converges to a mixture of Poisson laws, the mixing measure being the probability distribution of \( \lambda = \sum_{a} Z_{a,n} \).

References

Acknowledgements. This conference poster has been prepared using the postprint-1.1 archive from Andrew Fitzgibbon webpage. Random graphs in Fig. 1 and Fig. 2 have been produced using the R package SNA from Carter Butts.