Erdos-Renyi random graphs

May 11, 2011

1 Introduction

We denote by $\mathcal{G}_{ER}(n,p)$ the Erdos-Renyi random graph having set of vertices $V_n = \{1,\dots,n\}$ and set of edges $\mathcal{E}_{ER}(n,p)$ probabilistically constructed as follows. We have a family of independent random variables, indicized by all distinct (unordered) pairs of vertices $\{i,j\} \subseteq V$, $X_{\{i,j\}}$ all having identical distribution Ber(p) (Bernoullian of parameter p) and we put

$$\mathcal{E}_{ER}(n,p) = \{\{i,j\} \,|\, X_{\{i,j\}} = 1\}$$

2 The branching process approximation

Let $\mathcal{C}_1$ be the set of vertices in the connected component of $\mathcal{G}_{ER}(n,p)$ containing the vertex 1. We can imagine to construct $\mathcal{C}_1$ through a sort of cluster growth or epidemic spreading: we start from the node 1 and, at time 1 we consider its sons (infected nodes) as the nodes which are neighbors of 1, after we consider the nodes which are neighbors of the neighbors of 1 and so on. Formally, we consider the splitting $V_n = S_t \cup I_t \cup R_t$ to be interpreted, respectively, as the set of susceptible nodes, infected nodes, removed nodes, at time t and which are determined through an iterative process. Initially we set $I_0 = \{1\}$, $S_0 = V_n \setminus I_0$, and $R_0 = \emptyset$. Given the splitting at time t, we put

$$I_{t+1} = \{i \in S_t \,|\, X_{\{j,i\}} = 1 \text{ for some } j \in I_t\}\,,\ S_{t+1} = S_t \setminus I_{t+1}\,,\ R_{t+1} = R_t \cup I_t$$

Clearly, $\mathcal{C}_1 = \cup_{t\geq 0} I_t$ and, since the I_t are pairwise disjoint,

$$|\mathcal{C}_1| = \sum_{t=0}^{+\infty} |I_t| = \lim_{t\to+\infty} |R_t|$$

The branching process approximation works as follows. We define a family of independent auxiliary Ber(p) random variables $Y^t_{\{i,j\}}$ where $i,j \in \mathbb{N}$, $t \geq 0$ independent also from the $X_{\{i,j\}}$ and we define the process Z_t iteratively as

follows

$$Z_{t+1} = \begin{cases} 0 & \text{if } Z_t = 0 \\ \sum_{i\in I_t, j\in S_t} X_{\{i,j\}} + \sum_{i\in I_t, j\in S_t^c} Y^t_{\{i,j\}} \sum_{i=n+1}^{n+Z_t-|I_t|} \sum_{j\in V_n} Y^t_{\{i,j\}} & \text{if } Z_t > 0 \end{cases} \qquad Z_0 = 1 \tag{1}$$

Notice first of all that Z_t is indeed a branching process with binomial $B(n,p)$ offspring distribution. What does it have to do with our connected component? Let us try to explain the meaning of the three summation terms in (1). The first summation takes into account the nodes which become infected at time $t+1$, namely those which are connected by a path of length $t+1$ to the root node 1; however, by the way this terms is defined it counts more than one time those nodes which are neighbors of more than one node in I_t. Therefore we have an inequality $|I_{t+1}| \leq \sum_{i\in I_t, j\in S_t} X_{\{i,j\}}$ and therefore we also have $Z_t \geq |I_t|$. As a simple consequence we have that if Z_t dies, I_t becomes 0 at a certain point and the connected component is thus finite. Using the theory of branching process, we thus have the following result

Proposition 1. *Suppose that $p = \lambda/n$ with $\lambda < 1$. Then $\mathcal{C}_1$ is bounded in n with probability* 1.

For a better understanding of the quality of this approximation and also to get results for $\lambda > 1$, we have to study more carefully the branching process (1). First define $C_{t+1} = \sum_{i\in I_t, j\in S_t} X_{\{i,j\}} - |I_{t+1}|$ the collisions at time t (multiplicity due to vertices reached more than one time). The second summation is denoted by B_{t+1}, and represents the extra birth due to 'fake' vertices in S_t^c (already visited in the cluster growth): this term is needed in order to maintain the same offspring distribution at all times. The third term represents instead the offspring from vertices which are not in I_t, vertices which have made their appearance in the past due to the two phenomena of collisions and extra birth. It is convenient to introduce the bi-indicized sequence A_{st} defined by $A_{ss} = C_s + B_s$ (new immigrants at time s) while for $s < t$, A_{st} is the number of sons at time t due to immigrants born in the past at time s. Therefore, we have that

$$Z_t = I_t + \sum_{s=1}^{t} A_{st} \tag{2}$$

Hence, if we can estimate this terms A_{st} we will also be able to study how much Z_t and $|I_t|$ differ from each other.

The following lemma establishes a number of useful results:

Lemma 2. *Suppose that $p = \lambda/n$. The following facts are true*

(a) $\mathbb{E}(A_{st}) = \lambda^{t-s}\mathbb{E}[B_s + C_s]$

(b) $\mathbb{E}[B_{t+1}] \leq \frac{\lambda}{n}\mathbb{E}[Z_t \sum_{s=0}^{t} Z_s]$

(c) $\mathbb{E}[C_{t+1}] \leq \frac{\lambda^2}{n}\mathbb{E}[Z_t^2]$

Proof (Only a sketch to be completed by the reader).

(a) is a trivial consequence of the properties of branching processes.

(b) Consider $\mathcal{F}_t$ the σ-algebra generated by all the events in the cluster growth up to time t. It holds (explain why)

$$\mathbb{E}[B_{t+1}|\mathcal{F}_t] = \frac{\lambda}{n}|I_t|(|I_t| + |R_t|)$$

Prove now the thesis using the inequalities $|I_t| \leq Z_t$.

(c) It holds (explain why)

$$C_{t+1} \leq |\{(x, x', y) \in I_t^2 \times S_t \,|\, x < x',\, X_{\{x,y\}} = X_{\{x',y\}} = 1\}|$$

Now estimate as in (b), first considering $\mathbb{E}[C_{t+1}|\mathcal{F}_t]$ and then using the inequalities $|I_t| \leq Z_t$ and $|S_t| \leq n$. ■

The above lemma reduce the problem to the computation of second order moments of the process Z_t. We have the following results which are left without proof inviting the reader to try to prove it himself or rather looking for proofs in the literature on branching processes.

Lemma 3. *Suppose that $s > r$. Then,*

$$\mathbb{E}[Z_s Z_r] = \lambda^{s-r}\mathbb{E}[Z_r^2]$$

Lemma 4. *It holds*

$$\mathbb{E}[Z_{s+1}^2] = \lambda^2\mathbb{E}[Z_s^2] + \sigma^2\mathbb{E}[Z_s]$$

where $\sigma^2 = \lambda(1 - \lambda/n)$.

Corollary 5. *Suppose that $\lambda < 1$. There exists a positive constant C such that, for every n it holds*

$$\sum_{t=1}^{+\infty} \mathbb{E}(Z_t - |I_t|) \leq \frac{C}{n}$$

Proof (Sketch) Using (2), Lemmas 2, 3 prove first that

$$\sum_{t=1}^{+\infty} \mathbb{E}(Z_t - |I_t|) \leq \frac{\lambda}{n} \sum_{s=0}^{+\infty} \sum_{r=0}^{s-1} \lambda^{s-r}\mathbb{E}[Z_r^2] + \frac{\lambda + \lambda^2}{n} \sum_{s=0}^{+\infty} \mathbb{E}[Z_s^2]$$

Using now Lemma 4 with the fact that $\sigma^2 \leq \lambda < 1$, show that

$$\mathbb{E}[Z_s^2] \leq \frac{\lambda^s}{1 - \lambda}$$

Finally combine with the derivation above to get the result. ■

Corollary 6. *Suppose that $\lambda > 1$. There exists a positive constant C such that, for every n it holds*

$$\mathbb{E}(Z_t - |I_t|) \leq \frac{C}{n}\lambda^{2t} \tag{3}$$

Proof Exercise using arguments similar to the proof of Corollary 6. ■

Consider (2) for $t = a \ln n / \ln \lambda$: we obtain

$$\mathbb{E}(Z_t - |I_t|) \leq cn^{2a-1}$$

Therefore, if $a < 1/2$, we have that $\mathbb{E}(Z_t - |I_t|) \to 0$ for $n \to +\infty$. Instead, if $a \in [1/2, 1[$, since $n^{2a-1} << \lambda^t = \mathbb{E}[Z_t]$, we still have that $\mathbb{E}(Z_t - |I_t|) << \mathbb{E}[Z_t]$.

3 The connectivity threshold

Theorem 7. *Let p_n be a sequence in $[0, 1]$.*

(a) If $p_n = \frac{\ln n + \omega_n}{n}$ with $\omega_n \to +\infty$ then, for $n \to +\infty$,

$$\mathbb{P}(\mathcal{G}_{ER}(n, p_n) \text{ is connected}) \to 1$$

(b) If $p_n = \frac{\ln n - \omega_n}{n}$ with $\omega_n \to +\infty$ then, for $n \to +\infty$,

$$\mathbb{P}(\mathcal{G}_{ER}(n, p_n) \text{ is connected}) \to 0$$

Proof Let N_k be the number of connected components of cardinality exactly k inside $\mathcal{G}_{ER}(n, p_n)$ and notice that

$$\{\mathcal{G}_{ER}(n, p_n) \text{ is disconnected}\} \subseteq \{\sum_{k=1}^{\lfloor n/2 \rfloor} N_k \geq 1\} \tag{4}$$

$$\{\mathcal{G}_{ER}(n, p_n) \text{ is connected}\} \subseteq \{N_1 = 0\} \tag{5}$$

Condition (c2) and Markov inequality will lead to prove (a). Instead (c1) and a second order argument (Chebyschev) will lead to prove (b).

We start with (a). For every subset $H \subseteq V_n$ let E_H be the Bernoulli variable which is 1 iff H is isolated from the rest in $\mathcal{G}_{ER}(n, p_n)\}$ Clearly, $N_k \leq \sum_{H:|H|=k} E_H$ (why is not an equality?). Therefore

$$\mathbb{E}[N_k] = \sum_{H:|H|=k} \mathbb{E}[E_H] = \binom{n}{k}(1 - p_n)^{k(n-k)}$$

It follows that

$$\mathbb{P}\left(\sum_{k=1}^{\lfloor n/2 \rfloor} N_k \geq 1\right) \leq \sum_{k=1}^{\lfloor n/2 \rfloor} \mathbb{E}[N_k] \leq \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k}(1 - p_n)^{k(n-k)}$$

To estimate the last summation above we proceed as follows. We first consider an integer sequence $\alpha_n \leq n/2$ to be determined later and we split

$$\sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k} (1-p_n)^{k(n-k)} = \sum_{k=1}^{\alpha_n} \binom{n}{k} (1-p_n)^{k(n-k)} + \sum_{k=\alpha_n+1}^{\lfloor n/2 \rfloor} \binom{n}{k} (1-p_n)^{k(n-k)}$$

Using the classical inequalities

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k \ , \ \ln(1+x) \leq x$$

we obtain (check this)

$$\sum_{k=1}^{\alpha_n} \binom{n}{k} (1-p_n)^{k(n-k)} \leq \sum_{k=1}^{\alpha_n} \left[k^{-k} e^k e^{2k^2 \frac{\ln n}{n}}\right] \left[e^{-k\omega_n\left(1-\frac{k}{n}\right)}\right]$$

To prove convergence to 0 we now show that there exists a summable sequence a_k such that

$$k^{-k} e^k e^{2k^2 \frac{\ln n}{n}} \leq a_k \ \ \forall k, n$$

and moreover we show that if $\alpha_n \sim n^{3/4}$ for $n \to +\infty$,

$$\sup_{k \leq \alpha_n} \left[e^{-k\omega_n\left(1-\frac{k}{n}\right)}\right] \to 0$$

for $n \to +\infty$. This proves that $\sum_{k=1}^{\alpha_n} \binom{n}{k} (1-p_n)^{k(n-k)}$ is indeed infinitesimal for $n \to +\infty$ (the reader should check all these facts). Regarding the second term, check the following steps

$$\sum_{k=\alpha_n+1}^{\lfloor n/2 \rfloor} \binom{n}{k} (1-p_n)^{k(n-k)} \leq \sum_{k=\alpha_n+1}^{\lfloor n/2 \rfloor} \left(\frac{en^{1/2}}{\alpha_n}\right)^k \leq \frac{n}{2} \left(\frac{en^{1/2}}{\alpha_n}\right)^{\alpha_n}$$

Conclude that this term is also infinitesimal if $\alpha_n \sim n^{3/4}$ for $n \to +\infty$. This proves (a).

For (b) we use a second order method.

$$\mathbb{P}(N_1 = 0) \leq \mathbb{P}(|N_1 - \mathbb{E}[N_1]|^2 \geq \mathbb{E}[N_1])^2) \leq \frac{\mathrm{Var}(N_1 =}{\mathbb{E}[N_1])^2}$$

Now we can compute as follows (check this)

$$\mathbb{E}[N_1]) = n(1-p_n)^{n-1} \ , \ \ \mathrm{Var}(N_1) = n(n-1)(1-p_n)^{2n-3} - n^2(1-p_n)^{2n-2} + n(1-p_n)^{n-1}$$

Finally, insert these expressions in the estimation above and prove convergence to 0. ■