Graph mining
SD212
Graph models

Thomas Bonald

2018 – 2019

These lecture notes present various graph models. These are very useful to generate random instances of graphs and thus assess the efficiency of graph algorithms. We focus on undirected graphs but most results easily extend to directed graphs. We refer the reader to [1, 2] for further information on random graphs.

1 Erdös-Rényi graphs

An Erdös-Rényi graph is characterized by two parameters: the number of nodes n and the probability p that any node pair is connected. The degree of each node has a Binomial distribution with parameters $n-1, p$. When $n \to +\infty$ and $p \to 0$ with $np \to \lambda$, this tends to a Poisson distribution with parameter λ. In this limiting regime, cycles are extremely rare so that the local structure around each node is that of a tree. Specifically, it is a Galton-Watson tree whose offspring distribution is Poisson with parameter λ (see the appendix). We denote by q_λ the probability that such a tree is finite. There are three regimes:

- **Subcritical regime** ($\lambda < 1$)**:** The Galton-Watson tree is a.s. finite (that is, $q_\lambda = 1$). This means that the size of each connected component of the graph is in $O(1)$ (that is, is negligible compared to n). In particular, the number of connected components of the graph is in $O(n)$.
- **Supercritical regime** ($\lambda > 1$)**:** There is a positive probability $p_\lambda = 1 - q_\lambda$ that the tree is infinite. Since there are only n nodes, this implies the existence of a *giant component* containing a fraction p_λ of the nodes, with $p_\lambda \to 1$ when $\lambda \to +\infty$. The other nodes belong to the $O(n)$ other connected components, each of size $O(1)$. The size of these connected components tends to decrease with λ by the duality principle (see Theorem 1).
- **Critical regime** ($\lambda = 1$)**:** It can be shown that the largest connected component contains $O(n^{2/3})$ nodes (that is, a negligible fraction of the nodes when n is large). The rest of the graph behaves as in the subcritical case.

An interesting question in the supercritical regime is that of the value of λ (as a function of n) beyond which the graph is connected with high probability. It turns out that the graph is connected with high probability whenever there are no isolated nodes with high probability. Since each node is isolated with probability $e^{-\lambda}$, the number of isolated nodes has a binomial distribution with parameter $n, e^{-\lambda}$, which is close to a Poisson distribution with parameter $ne^{-\lambda}$ (because the probability $e^{-\lambda}$ of a node to be isolated must be low). Thus the probability of having at least one isolated node is approximately $1 - e^{-ne^{-\lambda}}$. We want this probability to be low, say lower than some ϵ, so that the graph is connected with high probability. This implies that $\epsilon \approx ne^{-\lambda}$. We deduce that:

$$\lambda \approx \ln n - \ln \epsilon.$$

Thus the average degree must be of the order of $\ln n$. For $\epsilon = 0.05$, the degree must be at least 10 for a graph of $n = 1000$ nodes, that is $p > 0.01$.

2 Preferential attachment

The empirical degree distribution of nodes in a large Erdös-Rényi graph is close to a Poisson distribution, while most real graphs have a power-law degree distribution, that is $p_k \sim 1/k^\alpha$ for some $\alpha > 1$. Barabasi and Albert have proposed a model in 1999 that both explains this suprising property of real graphs and allows one to generate random instances of graphs with such a distribution [1]. The idea is to start from some initial small graph and to let this graph grow by adding nodes that are connected to existing nodes in proportion to their degrees, a phenomenon called *preferential attachment*. A similar model has been proposed by Yule in 1925 for the phylogenetic tree (random branching of species) [3].

The Barabasi-Albert has two parameters: the number of nodes n and the degree of new nodes d, with $d < n$. The initial graph is typically a clique of d nodes (but the principle applies to any non-empty initial graph). At time $t = 1, \ldots, n - d$, a new node of degree d is added. Its d neighbors are chosen at random among the $d + t - 1$ existing nodes with probabilities proportional to their degrees. Observe that the first added node is connected to the d initial nodes so that the graph at time $t = 1$ is a clique of $d + 1$ nodes. It can be shown that, for large n, the degree distribution of the final graph has a power law with parameter $\alpha = 3$.

3 Configuration model

The configuration model is useful to generate a graph with a specific sequence of node degrees $d_1, \ldots, d_n$. Clearly, not all sequences are allowed since the total degree $d_1 + \ldots + d_n$ must be even (it is equal to $2m$ where m is the number of edges). If you allow *multi-graphs* (that is, graphs with possibly loops and multi-edges), then this condition is sufficient (edges can be added sequentially between two arbitrary nodes i, j whose current degrees are less than d_i, d_j, respectively). Now if you want a *simple* graph (that is a graph without loops nor multi-edges), not all sequences of even sum are allowed. For instance, the sequence $d_1 = 3, d_2 = 3, d_3 = 3, d_4 = 1$ is not allowed (you need to connect nodes 1,2,3 to all other nodes but the degree of node 4 must be 1). A sequence of degrees that corresponds to a simple graph is called *graphical*.

Havel-Hakimi algorithm. There is a simple iterative algorithm to check whether a sequence is graphical. First order the sequence so that $d_1 \geq d_2 \geq \ldots \geq d_n$ and check that $d_1 \leq n - 1$. Then remove 1 unit from the degree of the d_1 nodes following node 1, so that the sequence becomes:

$$0, d_2 - 1, d_3 - 1, \ldots, d_{k+1} - 1, d_{k+2}, \ldots, d_n,$$

where $k = d_1$. Iterate this process until the sequence is $0, \ldots, 0$. If this is not possible (that is, negative values appear), the sequence is not graphical. It is for instance easy to check that the sequence $3, 3, 2, 2$ is graphical, unlike the sequence $3, 3, 3, 1$.

This is known as the Havel-Hakimi algorithm. Observe that this is a constructive proof, in the sense that it provides a simple graph with the target degree sequence. In the first step of the algorithm, node 1 is connected to nodes $2, \ldots, d_1 + 1$. There are no loops nor multi-edges. Moreover, node 1 has degree d_1 and will not be revisited in the rest of the algorithm, so that no multi-edges are created.

Random configuration. Now assume that the sequence $d_1, \ldots, d_n$ is graphical (e.g., extracted from a real graph). The Havel-Hakimi algorithm provides one instance of a simple graph with this degree sequence. But the resulting graph is very specific, with nodes of highest degrees connected between them. Moreover, we would like to generate many random instances of graphs with this degree sequence.

The configuration model consists in connecting at random the $d_1, \ldots, d_n$ half-edges of the nodes. Specifically, the first half edge of node 1 is connected to one of the $2m - 1$ other half-edges, chosen uniformly at random. The next half-edge of node 1 (or node 2 if node 1 has no more free half-edge) is then connected to one of the $2m - 3$ other free half-edges, and so on. The resulting shuffling of the half-edges is called a

configuration, hence the name of the model. The total number of configurations is:

$$(2m-1)!! = \prod_{k=1}^{m}(2k-1).$$

Any random matching of half-edges as described above leads to one of these configurations, chosen uniformly at random (the order in which nodes are considered does not matter). The problem is that many such configurations correspond to multi-graphs.

Occurence of multi-graphs. We would like to estimate the probability that a random configuration is a multi-graph, under the assumption that $d_1, \dots, d_n$ are much smaller than $2m$. Let D be a random variable having the empirical degree distribution, that is

$$\Pr(D = k) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{d_i = k\}}.$$

Observe that

$$\mathrm{E}(D) = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{2m}{n}.$$

We shall see that the occurence of loops and multi-edges depends on the parameter:

$$\gamma = \frac{\mathrm{E}(D(D-1))}{\mathrm{E}(D)} = \frac{\mathrm{E}(D^2)}{\mathrm{E}(D)} - 1$$

First observe that there are $(2m-3)!!$ configurations with a given matching between two half-edges. In particular, the total number of configurations with a loop at node i is at most

$$\binom{d_i}{2}(2m-3)!!$$

since configurations with more than one loop at node i are counted several times in the above expression. We deduce that the probability of a loop at node i is upper bounded by

$$\binom{d_i}{2}\frac{(2m-3)!!}{(2m-1)!!} = \frac{d_i(d_i-1)}{2(2m-1)}.$$

Thus the total number of loops L satisfies

$$\mathrm{E}(L) \le \frac{\sum_{i=1}^{n} d_i(d_i-1)}{2(2m-1)} = \frac{n\mathrm{E}(D(D-1))}{2(2m-1)} = \gamma\frac{m}{2m-1} \approx \frac{\gamma}{2}.$$

Moreover, provided $d_1^2, \dots, d_n^2$ are much smaller than m, loops are rare and approximately independent events and L is upper bounded with high probability by a Poisson random variable with parameter $\gamma/2$.

Similarly, there are $(2m-5)!!$ configurations with a given matching between two half-edges of node i and two half-edges of j. Thus the total number of configurations with a multi-edge between i and j is at most

$$2\binom{d_i}{2}\binom{d_j}{2}(2m-5)!!$$

since configurations with more than two edges between nodes i and j are counted several times in the above expression. We deduce that the probability of a multi-edge between nodes i and j is upper bounded by

$$2\binom{d_i}{2}\binom{d_j}{2}\frac{(2m-5)!!}{(2m-1)!!} = \frac{d_i(d_i-1)d_j(d_j-1)}{2(2m-1)(2m-3)}.$$

Thus the total number of multi-edges M satisfies

$$\mathrm{E}(M) \leq \sum_{i<j} \frac{d_i(d_i-1)d_j(d_j-1)}{2(2m-1)(2m-3)}.$$

Using the fact that

$$\begin{aligned}\sum_{i<j} d_i(d_i-1)d_j(d_j-1) &= \frac{1}{2}\sum_{i\neq j} d_i(d_i-1)d_j(d_j-1),\\ &\leq \frac{1}{2}\sum_{i,j} d_i(d_i-1)d_j(d_j-1),\\ &= \frac{n^2}{2}\mathrm{E}(D(D-1))^2,\\ &= 2m^2\gamma^2,\end{aligned}$$

we obtain

$$\mathrm{E}(M) \leq \frac{m^2}{(2m-1)(2m-3)}\gamma^2 \approx \frac{\gamma^2}{4}.$$

Moreover, provided $d_1^2, \ldots, d_n^2$ are much smaller than m, multi-edges are rare and approximately independent events and M is upper bounded with high probability by a Poisson random variable with parameter $\gamma^2/4$.

Appendix

Galton-Watson trees

A Galton-Watson tree is defined recursively, starting from any given node (the root of the tree), by the property that the number of children of each node is a random number taken from some fixed distribution. We here focus on a Poisson distribution with parameter λ. Observe that the tree may be finite or infinite.

Let Z_k be the number of nodes at the k-th generation, with $Z_0 = 1$. Then Z_{k+1} has the same distribution as

$$\sum_{i=1}^{Z_k} X_i,$$

where $X_1, X_2, ...$ are i.i.d. random variables of Poisson distribution with parameter λ. Observe in particular that Z_1 (the number of children of the root node) has a Poisson distribution with parameter λ.

We denote by q_λ the probability that the tree is finite, i.e., $Z_k = 0$ for some $k \geq 1$. We refer to q_λ as the extinction probability.

Proposition 1 *The extinction probability is the smallest solution q_λ over $[0,1]$ of the equation:*

$$q = e^{\lambda(q-1)}. \tag{1}$$

In particular, $q_\lambda = 1$ if and only if $\lambda \leq 1$.

Proof. The proof relies on the fact that, denoting by G_k the generating function of Z_k,

$$G_{k+1}(t) = G_k(G(t)),$$

where $G(t) = e^{\lambda(t-1)}$ is the generating function of a Poisson random variable with parameter λ. In particular,

$$G_{k+1}(t) = \underbrace{G \circ \ldots \circ G}_{k+1}(t) = G(G_k(t)).$$

Observing that $G_k(0)$ is the probability that $Z_k = 0$, we get by taking the limit,

$$q_\lambda = e^{\lambda(q_\lambda - 1)}.$$

□

Using the fact that

$$\mathrm{E}(Z_{k+1}) = \mathrm{E}(Z_k)\mathrm{E}(X_1) = \lambda \mathrm{E}(Z_k),$$

we get

$$\mathrm{E}(Z_k) = \lambda^k.$$

If $\lambda < 1$, the total number of nodes has mean

$$\sum_{k \geq 0} \mathrm{E}(Z_k) = \frac{1}{1-\lambda}.$$

If $\lambda \geq 1$, the total number of nodes has infinite mean (because the tree is infinite with positive probability). A natural question is that of the mean (and distribution) of the number of nodes *conditioned* on extinction. It turns out that it can be completely characterized in terms of another Galton-Watson tree:

Theorem 1 (Duality principle) *Let $\lambda \geq 1$. The Galton-Watson tree with parameter λ conditioned on extinction has the same distribution as a Galton-Watson tree with parameter λq_λ.*

Proof. To prove the result, we describe the tree as a sequence of random variables $X_1, X_2, \ldots$ corresponding to the number of children in the breadth-first search exploration of the tree. The total number of children (excluding the parents) after k steps is then

$$S_k = \sum_{i=1}^{k} X_i - k + 1.$$

In particular, there is extinction after k steps if $S_1, \ldots, S_{k-1} > 0$ and $S_k = 0$.

Let $x_1, \ldots, x_k$ be any sequence of random variables corresponding to extinction after k steps. Denoting by A the event of extinction, we have

$$\begin{aligned} \mathrm{P}(X_1 = x_1, \ldots, X_k = x_k | A) &= \frac{\mathrm{P}(X_1 = x_1, \ldots, X_k = x_k)}{\mathrm{P}(A)}, \\ &= \frac{1}{q_\lambda} e^{-k\lambda} \frac{\lambda^{x_1 + \ldots + x_k}}{x_1! \ldots, x_k!}, \end{aligned}$$

Now considering the Galton-Watson tree with parameter λq_λ, we have

$$\mathrm{P}(X'_1 = x_1, \ldots, X'_k = x_k) = e^{-k\lambda q_\lambda} \frac{(\lambda q_\lambda)^{x_1 + \ldots + x_k}}{x_1! \ldots, x_k!},$$

Using the fact that $x_1 + \ldots + x_k = k - 1$, we get the ratio:

$$\frac{\mathrm{P}(X_1 = x_1, \ldots, X_k = x_k | A)}{\mathrm{P}(X'_1 = x_1, \ldots, X'_k = x_k)} = \frac{1}{q_\lambda} e^{k\lambda(q_\lambda - 1)} \frac{1}{q_\lambda^{k-1}} = \left(\frac{e^{\lambda(q_\lambda - 1)}}{q_\lambda} \right)^k = 1,$$

where the last equality follows from (1). □

The following result shows that when λ increases from 1 to $+\infty$, the extinction probability q_λ decreases; more suprisingly, the typical size of the tree *conditionned on extinction* λq_λ also decreases.

Proposition 2 *Both q_λ and λq_λ are decreasing functions of λ over $(1, +\infty)$.*

Proof. Let f be the function defined over $[0, 1]$ by $f(q) = e^{\lambda(q-1)}$, with $\lambda > 1$. We have $f(1/\lambda) = e^{1-\lambda} < 1/\lambda$ (because $e^{\lambda-1} > \lambda$). This shows that $q_\lambda < 1/\lambda$, that is $\lambda q_\lambda < 1$.

Viewing q_λ as a function of λ, we have

$$\ln q_\lambda = \lambda(q_\lambda - 1)$$

so that

$$\frac{q'_\lambda}{q_\lambda} = q_\lambda - 1 + \lambda q'_\lambda,$$

and

$$q'_\lambda = \frac{q_\lambda(q_\lambda - 1)}{1 - \lambda q_\lambda} < 0.$$

This show that q_λ is a decreasing function of λ.

Finally,

$$(\ln(\lambda q_\lambda))' = \frac{1}{\lambda} + \frac{q'_\lambda}{q_\lambda} = \frac{1}{\lambda} + \frac{q_\lambda - 1}{1 - \lambda q_\lambda} = \frac{1 - \lambda}{1 - \lambda q_\lambda} < 0.$$

Thus λq_λ is a decreasing function of λ. □

References

[1] A.-L. Barabási. *Network science*. Cambridge University Press, 2016.

[2] R. Van Der Hofstad. *Random graphs and complex networks*. Cambridge University Press, 2017.

[3] G. U. Yule. A mathematical theory of evolution, based on the conclusions of dr. jc willis, frs. *Philosophical Transactions of the Royal Society of London Series B*, 213:21–87, 1925.