Seminar über Algorithmen 28.11.2017

Random Graphs

Foundations of Data Science, 8.1-8.2

Benjamin Berendsohn *Wolfgang Mulzer*

Note: I will write $x_n \to c$ for $x_n \overset{n\to\infty}{\longrightarrow} c$ and $x_n \sim y_n$ for $\frac{x_n}{y_n} \to 1$.

1 Terminology

Definition 1. Let $n \in \mathbb{N}$, $p \in (0,1)$. The Erdös-Rényi Graph $G(n,p)$ is a random graph where every possible edge is present with probability p (independent of other edges).

$G(n,p)$ is a random variable. We usually look at families of random graphs where $n \to \infty$ and p is a function of n. We say a property of such a family occurs *almost surely* if its probability converges to 1 as n approaches infinity.

2 General Observations

Even though the edges are chosen independently, certain global properties emerge for specific choices of p. Specifically, for some properties, there is a *threshold* p', meaning if p is sufficiently smaller than p', the property occurs with high (low) probability; and if p is sufficiently larger then p', then the property occurs with low (high) probability. We will look at several properties with different thresholds.

3 Degree distribution

Example 1. In $G(n,1/2)$, for $\varepsilon > 0$, the degree of each vertex is almost surely within $(1 \pm \varepsilon)n/2$.

Proof notes. The standard deviation is $\sigma = \sqrt{\text{Var}} = \frac{\sqrt{n-1}}{2}$. By Chebychev's inequality, we have $\Pr[|x - E[X]| \geq c\sigma] \leq 1/c^2$. Therefore, any constant fraction of the probability mass lies within $(n-1)/4 \pm c\sqrt{n-1}/2$, or within $(n-1)/4 \pm \varepsilon$ for large n. □

For $G(n,p)$, the degree distribution is also a binomial distribution:

$$\Pr[\text{Vertex has degree } k] = \binom{n-1}{k} p^k (1-p)^{n-k-1}$$
$$= \frac{n-k}{n(1-p)} \binom{n}{k} p^k (1-p)^{n-k} \sim \binom{n}{k} p^k (1-p)^{n-k}$$

for $k \in o(n)$ and $p \to 0$.

Vertex degrees in real-world graphs often have a power law distribution, where the probability of a degree k decreases polynomially in k, not exponentially as with a binomial distribution. We won't consider this here.

Theorem 1 (Theorem 8.1, Corollary 8.2 without approximations)**.** *Let $\varepsilon \in (0,1)$. If $p \geq 9\frac{\ln n}{(n-1)\varepsilon^2}$, then almost surely every vertex has degree in the range $(1 \pm \varepsilon)(n-1)p$.*

Proof. The degree $\deg(v)$ is a sum of $m = n-1$ independent Bernoulli random variables $x_1, x_2, \ldots, x_m$, where x_i is the indicator variable that the i^{th} edge from v is present. Applying the Chernov Bound (Theorem 12.6):

$$\Pr[|mp - \deg(v)| \geq \varepsilon mp] \leq 3e^{-mp\varepsilon^2/8} \leq 3n^{-9/8} \to 0.$$

□

4 The binomial distribution approaches the poisson distribution

Lemma 1. *Assume $k \leq o(n)$ and $\binom{n}{k} \sim n^k/k!$. As n goes to infinity, the binomial distribution with d/n*

$$\Pr[k] = \binom{n}{k} \left(\frac{d}{n}\right)^k \left(1 - \frac{d}{n}\right)^{n-k}$$

approaches the poisson distribution

$$\Pr[k] = \frac{d^k}{k!} e^{-d}.$$

Therefore, in $G(n, d/n)$ for a constant d, the degree distribution is a poisson distribution.

Proof sketch. Note that $(1-d/n)^n \sim e^{-d}$. It remains to show that $(1-d/n)^k \to 1$. Clearly, $(1-d/n)^k < 1$. By Bernoulli's inequality, for $n \geq d$,

$$\left(1 - \frac{d}{n}\right) \geq 1 - \frac{kd}{n} \to 0.$$

□

The authors seem to claim that $\binom{n}{k} \sim n^k/k!$ follows from $k \in o(n)$, which is not clear to me. It certainly is true if k is a constant.

Example 2. In $G(n, 1/n)$, there is a vertex of degree $\Omega(\log n / \log\log n)$ with high (constant) probability if $n = k!$ for some k.

Proof notes. As argued in the book, $\log n / \log\log n \leq k$.

Let A_i be the event that the vertex i has degree k and let A be the event that any vertex has degree k. Then

$$\Pr[A] = \Pr\left[\bigcup_{i=0}^{n} A_i\right] \geq \sum_{i=1}^{n} \Pr[A_i] - \sum_{i<j} \Pr[A_i \cap A_j]$$

by the Bonferroni inequality[1]. Let i, j be vertices, and B_{ij} the event that the edge $\{i, j\}$ exists. Then we can calculate

$$\begin{aligned}
&\Pr[A_i \mid B_{ij}] \sim \frac{k}{en},\\
&\Pr[A_i \mid \overline{B_{ij}}] \sim \frac{1}{en}, \text{ and therefore}\\
&\Pr[A_i \cap A_j] = \Pr[B_{ij}] \cdot \Pr[A_i \cap A_j \mid B_{ij}] + \Pr[\overline{B_{ij}}] \cdot \Pr[A_i \cap A_j \mid \overline{B_{ij}}]\\
&\sim \frac{1}{n} \cdot \left(\frac{k}{en}\right)^2 + \left(1 - \frac{1}{n}\right)\left(\frac{1}{en}\right)^2 \sim \frac{k^2}{e^2 n^3} + \frac{1}{e^2 n^2},
\end{aligned}$$

using independence of A_i and A_j under B_{ij} or $\overline{B_{ij}}$. Summing up, we get

$$Pr[A] \gtrsim n\frac{1}{en} - n^2\left(\frac{k^2}{e^2 n^3} + \frac{1}{e^2 n^2}\right) \sim \frac{1}{e} - \frac{1}{e^2}.$$

□

[1] `https://en.wikipedia.org/wiki/Boole%27s_inequality`

5 Existence of Triangles

Lemma 2. *The expected number of triangles in $G(n, d/n)$ converges to $d^3/6$ from below.*

Proof notes. More precisely, if x_n is the number of triangles,

$$\mathrm{E}[x_n] = \binom{n}{3} p^3 = \binom{n}{3}\left(\frac{d}{n}\right)^3 = \frac{d^3}{6} \cdot \frac{(n-1)(n-2)}{n^2}.$$

□

We now want to find the probability that $G(n, d/n)$ contains at least one triangle. From the lemma it not necessarily follows that this probability is high; a huge amount of triangles could be concentrated in a tiny number of graphs. But this would mean that the variance is high, which we will prove to be false.

Lemma 3. *Let x be the number of triangles in $G(n, d/n)$. Then* $\mathrm{Var}[x] \leq \mathrm{E}[x] + o(1)$.

Proof notes. $\binom{n}{3}p^3$ is exactly the expectation value of x, therefore part 3 contributes $\mathrm{E}[x]$ to the whole sum and

$$\mathrm{E}[x^2] \leq \mathrm{E}^2[x] + \mathrm{E}[x] + o(1),$$

yielding the slightly stronger upper bound $\mathrm{Var}[x] \leq \mathrm{E}[x] + o(1)$. □

Theorem 2. *Let x be the number of triangles in $G(n, d/n)$. Then*

$$\Pr[x = 0] \leq \frac{6}{d^3} + o(1).$$

Proof notes. By Chebychev's inequality,

$$\Pr[x = 0] \leq \frac{\mathrm{Var}[x]}{\mathrm{E}^2[x]} \leq \mathrm{E}[x] + \frac{o(1)}{\mathrm{E}[x]} = \mathrm{E}[x] + o(1).$$

□

Thus, for $d > \sqrt[3]{6}$, $\Pr[x = 0] < 1$ and $G(n, d/n)$ has a triangle with nonzero probability, and for $p \in \omega(1/n)$, $G(n, d/n)$ almost surely has a triangle.[2] Furthermore, if $p \in o(1/n)$, $\mathrm{E}[x] \to 0$, so $G(n, p)$ almost surely has no triangle.

[2] Note that d is not actually required to be a constant in the above proofs, so $p \in \omega(1/n)$ can be written as $d \in \omega(1)$.

Table 1: Phase transitions

Probability	**Transition**
$p < o(\frac{1}{n})$	Forest of trees No component of size greater then $\mathcal{O}(\log n)$
$p = \frac{d}{n}, d < 1$	Cycles appear No component of size greater then $\mathcal{O}(\log n)$
$p = \frac{1}{n}$	Components of size $\mathcal{O}\left(n^{2/3}\right)$
$p = \frac{d}{n}, d > 1$	Giant component plus $\mathcal{O}(\log n)$ components
$p = \frac{1}{2}\frac{\ln n}{n}$	Giant component plus isolated vertices
$p = \frac{\ln n}{n}$	**Disappearance of isolated vertices** **Appearance of Hamiton circuit** Diameter $\mathcal{O}(\log n)$
$p = \sqrt{\frac{2\ln n}{n}}$	**Diameter 2**
$p = \frac{1}{2}$	Clique of size $(2-\varepsilon)\ln n$

6 Phase Transitions

Besides the existence of triangles, there are other graph properties that arise more abruptly as p passes some threshold.

Definition 2. Given some property, if there exists a function $p(n)$ such that

- for all $p_1(n) \leq o(p(n))$, $G(n, p_1(n))$ almost surely does not have the property; and
- for all $p_2(n) \geq \omega(p(n))$, $G(n, p_2(n))$ almost surely has the property,

then we say that a *phase transition* occurs, and $p(n)$ is the *threshold*.

If for $c < 1$, $G(n, cp(n))$ almost surely does not have the property and for $c > 1$, $G(n, cp(n))$ almost surely has the property, then $p(n)$ is a *sharp threshold*.

From the discussion in section 5, we know that the existence of triangles has a threshold at $p = 1/n$. Note that for non-sharp thresholds, multiplicative constants do not matter, so we could write the threshold as $\Theta(1/n)$.

To estimate the probabilty of the existence of something, we often use a variable x that *counts* the number of occurrences of an item in a random graph and calculate expected value and variance of x. This lets us apply the following two theorems:

Theorem 3 (First moment method)**.** *Let x_n be a sequence of nonnegative integer random variables. If $E[x_n]$ goes to zero as n goes to infinity, $Pr[x_n = 0]$ goes to one, i.e. $x_n = 0$ almost surely is true.*

Therefore, if for some p the expected value of the item count goes to zero, the probability of the existence of an item also goes to zero.

Theorem 4 (Second moment method, Theorem 8.3)**.** *Let x_n be a sequence of random variables. Let there be an $\varepsilon > 0$ such that $\mathrm{E}[x_n] > \varepsilon$ for all n, i.e. $\liminf(x_n) > 0$. If $\mathrm{Var}[x_n] \leq o(\mathrm{E}^2[x_n])$, then x_n is almost surely greater than zero.*

Proof notes: The theorem in the book only requires $\mathrm{E}[x] > 0$, which is either incorrect or written in a weird way. The condition $\mathrm{E}[x_n] > 0$ for all n isn't sufficient, as $\mathrm{E}[x_n]$ could still converge to 0 and then $\mathrm{Var}[x]/\mathrm{E}^2[x]$ not necessarily converges to 0.

Note that $\mathrm{Var}[x_n] \leq o(\mathrm{E}[x_n])$ is equivalent to $\sigma[x_n] \leq o(\mathrm{E}[x_n])$, we require the standard deviation to be significantly smaller than the expectation.

Corollary 1 (Corollary 8.4)**.** *Let x_n be a sequence of random variables as above. If $\mathrm{E}[x^2] \leq \mathrm{E}^2[x](1 + o(1))$, then x is almost surely greater than zero.*

The second moment method can be used to show that for some p, the probability of the existence of an item goes to one. Note that computing the expected value is often much easier than computing the variance (respectively $\mathrm{E}[x^2]$).

7 Graph diameter two

A random graph having diameter less or equal two is an example of a property with sharp threshold.

Theorem 5 (Theorem 8.5)**.** *The property "$G(n, p)$ has diameter two" has a sharp threshold at $p = \sqrt{\frac{2\ln n}{n}}$.*

Proof. Note that $\binom{n}{2} \sim n^2/2$ and

$$(1 - c^2 \ln n/n)^n \sim e^{-c^2 \ln n}$$

□

8 Disappearance of Isolated Vertices

Theorem 6 (Theorem 8.6)**.** *The property "$G(n,p)$ has isolated vertices" has a sharp threshold at $\frac{\ln n}{n}$.*

Proof notes. With $p = c\frac{\ln n}{n}$, we get

$$\mathrm{E}[x] = \left(1 - \frac{c \ln n}{n}\right)^{n-1} \sim \left(1 - \frac{c \ln n}{n}\right)^{n}, \text{ as } \left(1 - \frac{c \ln n}{n}\right) \to 1.$$

□

9 Hamilton circuits

The existence of a hamilton circuit is a property with a threshold that can't be directly found with the first and second moment methods. At the value of p at which the expected *number* of hamilton circuits starts to go to infinity, these hamilton circuits are highly concentrated in a small number of graphs, i.e. the variance is high. This illustrates the necessity of the second moment method.

Theorem 7. *The expected number of hamilton circuits in $G(n, d/n)$ is 0 for $d \leq e$ and ∞ for $d > e$.*

Proof notes. Using Stirling's approximation:

$$\mathrm{E}[x] = \frac{1}{2}(n-1)!\left(\frac{d}{n}\right)^n = \frac{1}{2n}n!\left(\frac{d}{n}\right)^n$$
$$\sim \frac{1}{2n}\sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left(\frac{d}{n}\right)^n = \sqrt{\frac{\pi}{2n}}\left(\frac{d}{e}\right)^n$$

□

Still, the threshold for the *existence* of hamilton circuits can't be $p = e/n$, as such graphs are with high probability not even connected. The actual threshold is $\frac{1}{n}\ln n$ (not proved here).