CS271 Randomness & Computation **Fall 2011**

Lecture 15: October 13

Lecturer: Alistair Sinclair *Based on scribe notes by:*

B. Godfrey, J. Sanders; P. Sarathi Dey, A. Prakash

15.1 The Power of Two Choices

In the previous lecture, we saw that when one throws $m = n$ balls into n bins independently and uniformly at random, the maximum load (number of balls in any bin) is $\Theta(\ln n / \ln \ln n)$ with probability tending to 1 as $n \to \infty$.

Suppose now that instead of placing each ball in a single random bin, we choose d random bins for each ball, place it in the one that currently has fewest balls, and proceed in this manner sequentially for each ball. Clearly if $d = n$ the maximum load will be optimal: $\lceil m/n \rceil$. But this requires global knowledge of the state of all bins. Today we will show that with $d = 2$ choices, the maximum load is a.a.s. at most

$$\frac{\ln \ln n}{\ln 2} + \Theta(1)$$

That is, having only one more choice than the basic model reduces the maximum load exponentially.

This result has been discovered multiple times; see [Mi96] for more background and various extensions. The proof technique we use here is due to [ABKU].

15.1.1 Proof idea

Let B_i be the number of bins with load $\geq i$ at the end of the process. Suppose we could find upper bounds β_i so that $B_i \leq \beta_i$ w.h.p. Then

$$\Pr[\text{given ball is placed in a bin with load} \geq i] \quad \leq \quad \left(\frac{\beta_i}{n}\right)^2,$$

because both of the ball's choices must land in a bin with load $\geq i$. This gives us a crude upper bound on the number of bins with load $\geq i + 1$: the distribution of B_{i+1} is dominated by the binomial distribution $\text{Bin}\left(n, (\beta_i/n)^2\right)$. The mean of this distribution is β_i^2/n, so by a Chernoff bound we can take $\beta_{i+1} = c\beta_i^2/n$ for some constant c, and we have

$$\frac{\beta_{i+1}}{n} = c \cdot \left(\frac{\beta_i}{n}\right)^2.$$

Thus $\frac{\beta_i}{n}$ decreases quadratically, so for $i \approx \frac{\ln \ln n}{\ln 2}$ we will have $\beta_i < 1$, which implies that the maximum load is $\frac{\ln \ln n}{\ln 2}$ w.h.p.

15.1.2 Full proof

For algebraic convenience, set $\beta_6 = \frac{n}{2e}$. Note that $B_6 \leq \beta_6$ is trivial since there can be at most $\frac{n}{6} < \frac{n}{2e}$ bins with ≥ 6 balls in them. For $i > 6$, let

$$\beta_{i+1} = \frac{e\beta_i^2}{n}.$$

Now define the event $\mathcal{E}_i = \{B_i \leq \beta_i\}$ (so $\Pr[\mathcal{E}_6] = 1$). We have

$$\Pr[\neg\mathcal{E}_{i+1}|\mathcal{E}_i] \quad = \quad \Pr[B_{i+1} > \beta_{i+1}|\mathcal{E}_i] \quad \leq \quad \frac{\Pr[\text{Bin}(n, (\beta_i/n)^2) \geq \beta_{i+1}]}{\Pr[\mathcal{E}_i]}.$$

Note that the denominator is necessary here: we cannot claim that the numerator bounds the conditional probability, because once we condition on $\mathcal{E}_i$ the bin choices are no longer independent.

We now apply a Chernoff bound of the form $\Pr[X \geq e\mu] \leq e^{-\mu}$, which follows directly from a more general form discussed in Lecture 13. (Specifically, plug $\beta = e - 1$ into bound $(*)$ of Corollary 13.3.) Thus,

$$\Pr[\neg\mathcal{E}_{i+1}|\mathcal{E}_i] \quad \leq \quad \frac{e^{-\beta_i^2/n}}{\Pr[\mathcal{E}_i]} \quad \leq \quad \frac{1/n^2}{\Pr[\mathcal{E}_i]},$$

provided $\beta_i^2/n \geq 2\ln n$. Now to remove the conditioning we prove by induction on i that $\Pr[\neg\mathcal{E}_i] \leq \frac{i}{n^2}$. In the base case, $\Pr[\neg\mathcal{E}_6] = 0$. For the inductive step,

$$\Pr[\neg\mathcal{E}_{i+1}] = \Pr[\neg\mathcal{E}_{i+1}|\mathcal{E}_i]\Pr[\mathcal{E}_i] + \Pr[\neg\mathcal{E}_i] \leq \frac{1/n^2}{\Pr[\mathcal{E}_i]} \cdot \Pr[\mathcal{E}_i] + \frac{i}{n^2} \leq \frac{i+1}{n^2}.$$

So $\Pr[\neg\mathcal{E}_i] \leq i/n^2 \leq 1/n$ for all i such that $\beta_i^2/n \geq 2\ln n$.

Now let i^* be the minimum i for which $\beta_i^2 < 2n\ln n$. Then $i^* = \frac{\ln\ln n}{\ln 2} + O(1)$. (**Exercise:** prove this.) To get a feel for what's going on at this point, note that w.h.p. there are $\leq \sqrt{2n\ln n}$ bins with load $\geq i^*$, so the expected number of balls falling in bins with load $\geq i^* + 1$ is at most $2\ln n$. The following claim finishes the proof:

Claim 15.1 $\Pr[B_{i^*+2} \geq 1] \leq O\left(\frac{\log^2 n}{n}\right)$.

Proof: Define $\mathcal{E}_{i^*+1} = \{B_{i^*+1} \leq 6\ln n\}$. We have

$$\begin{aligned}
\Pr[\neg\mathcal{E}_{i^*+1}] &\leq \Pr[B_{i^*+1} \geq 6\ln n|\mathcal{E}_{i^*}] \cdot \Pr[\mathcal{E}_{i^*}] + \Pr[\neg\mathcal{E}_{i^*}] \\
&\leq \frac{\Pr\left[\text{Bin}(n, 2\ln n/n) \geq 6\ln n\right]}{\Pr[\mathcal{E}_{i^*}]} \cdot \Pr[\mathcal{E}_{i^*}] + \frac{1}{n} \\
&\leq \frac{1}{n^2} + \frac{1}{n} = O\left(\frac{1}{n}\right),
\end{aligned}$$

where the bound on $\Pr[\neg\mathcal{E}_{i^*}]$ comes from our previous calculation, and we have again used a Chernoff bound on the Binomial distribution. Finally,

$$\begin{aligned}
\Pr[B_{i^*+2} \geq 1] &\leq \Pr[B_{i^*+2} \geq 1 \,|\, \mathcal{E}_{i^*+1}] \cdot \Pr[\mathcal{E}_{i^*+1}] + \Pr[\neg\mathcal{E}_{i^*+1}] \\
&\leq \frac{\Pr\left[\text{Bin}\left(n, (6\ln n/n)^2\right) \geq 1\right]}{\Pr[\mathcal{E}_{i^*+1}]} \cdot \Pr[\mathcal{E}_{i^*+1}] + O\left(\frac{1}{n}\right) \\
&\leq \left(\frac{6\ln n}{n}\right)^2 \cdot n + O\left(\frac{1}{n}\right) = O\left(\frac{(\ln n)^2}{n}\right),
\end{aligned}$$

where we have used a simple union bound on the probability that the Binomial is nonzero. ■

Exercise 1 *Prove that the maximum load is* $\Omega(\ln \ln n)$ *w.h.p.*

Exercise 2 *Extend the above analysis to show that if each ball has* d *choices, the maximum load is* $\frac{\ln \ln n}{\ln d} + O(1)$ *w.h.p. Thus, additional choices beyond* 2 *affect only the constant.*

15.2 The Giant Component in $\mathcal{G}_{n,p}$

We consider our usual model of random graphs, $\mathcal{G}_{n,p}$, and look specifically at graphs where $p = \frac{c}{n}$, for some constant c.

Theorem 15.2 *For* $G \in \mathcal{G}_{n,p}$ *and constant* c *where* $p = \frac{c}{n}$,

- *For* $c < 1$, *then asymptotically almost surely the largest connected component of* G *is of size* $O(\log n)$.
- *For* $c > 1$, *then asymptotically almost surely there exists a single largest component of* G *of size* $\alpha n(1 + o(n))$, *where* α *is the unique solution in* $(0, 1)$ *to* $\alpha + e^{-\alpha c} = 1$. *Moreover, the next largest component in* G *has size* $O(\log n)$.

In the boundary case $c = 1$ things are more complicated. In this case the largest component is of size $O(n^{2/3})$, but we will not prove this here. Moreover, if $c = 1 + c'n^{-1/3}$ then the size of the largest component varies smoothly with c'. In other words, the "width" of the phase transition is $n^{-1/3}$.

Before proving this theorem, we look at the Galton-Watson branching process for a result that will help us in this endeavor.

15.2.1 Galton-Watson Branching Process

Let X be a random variable that takes non-negative, integer values. The branching process defined by X starts with a single node at time 0. At each subsequent time step, every node from the previous time step gives rise to a random number of children determined by X, independently of the other nodes. (See Figure 15.1.)

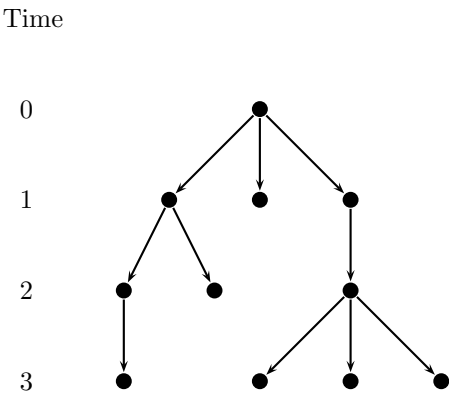

Figure 15.1: Galton-Watson Branching Process

The random variable Z_i counts the number of nodes at time i. By definition, $Z_0 = 1$; Z_i is distributed as Z_{i-1} independent copies of X. We can define the "extinction" of the process as the event when the number of nodes is zero. Thus,

$$\Pr[\text{ extinction }] = \lim_{n\to\infty} \Pr[Z_n = 0].$$

We will need the following basic fact about branching processes, found in [GS] or [Will].

Theorem 15.3 *For a branching process defined by a non-negative integer-valued r.v. X satisfying the conditions $\Pr[X = 1] < 1$ and $\Pr[X = 0] > 0$, we have:*

- *If $E[X] \le 1$ then $\lim_{n\to\infty} \Pr[Z_n = 0] = 1$, i.e., the process dies out a.a.s.*
- *If $E[X] > 1$ then $\lim_{n\to\infty} \Pr[Z_n = 0] = p^* < 1$, where p^* is the unique solution in $(0, 1)$ to $f(x) = x$, where $f(x)$ is the probability generating function*

$$f(x) = \sum_{i\ge 0} \Pr[X = i]x^i.$$

The conditions $\Pr[X = 1] < 1$ and $\Pr[X = 0] > 0$ rule out trivial extreme cases in which the theorem is actually false. In particular, if $\Pr[X = 1] = 1$ then $E[X] \le 1$ but clearly the process continues forever. And if $\Pr[X = 0] = 0$ then clearly the process continues forever with probability 1 regardless of the distribution of $E[X]$.

Before applying this result to the random graph scenario, we briefly sketch the proof of Theorem 15.3.

15.2.2 Proof of Theorem 15.3

Let f_n be the probability generating function of the random variable Z_n, *i.e.*,

$$f_n(x) = \sum_{i\ge 0} \Pr[Z_n = i]x^i$$

Note that $f_1(x) = f(x)$ (where $f(x)$ is the generating function of X, as in the theorem above) since Z_1 has the same distribution as X. For $n > 1$, Z_n is distributed as the sum of Z_1 many independent copies of Z_{n-1}. By the properties of generating functions we have $f_n(x) = f(f_{n-1}(x))$ for all $n > 1$. (Exercise: Verify this, e.g., by comparing coefficients.)

Let the probability of extinction at time n be $q_n := \Pr[Z_n = 0] = f_n(0)$. Then we have the following recursive relation:

$$q_n = f(q_{n-1}) \text{ for all } n \ge 1$$

where $q_0 = 0$. [This can also be proved directly as follows. If at time 1 the number of nodes is k, then there will be extinction at time n if and only if each of the k offspring gives rise to 0 children after $n - 1$ more levels. Consequently $q_n = \Pr[Z_n = 0] = \sum_{k\ge 0} \Pr[Z_1 = k] \Pr[Z_{n-1} = 0]^k = \sum_{k\ge 0} \Pr[X = k]q_{n-1}^k = f(q_{n-1})$.]

As the probability of extinction at time n is at least as large as the extinction probability at time $n - 1$, the sequence q_n is monotonically increasing with $0 < q_n \le 1$ for all n, *i.e.,*

$$0 < q_1 \le q_2 \le q_3 \le \cdots \le 1.$$

Since the sequence (q_i) is increasing and bounded, it must converge to a limit; thus as $n \to \infty$, $q_n \to q^*$ where $0 < q^* \le 1$. Also, the fact that f is continuous and $q_n = f(q_{n-1})$ for all $n \ge 1$ implies that q^* is a fixed point of the function $f(x)$, i.e., $q^* = f(q^*)$.

Observe that f is a strictly increasing convex function from $[0,1]$ to $[0,1]$ with $f(1) = 1$ and $f(0) > 0$. We have two different cases as depicted in Figure 15.2.

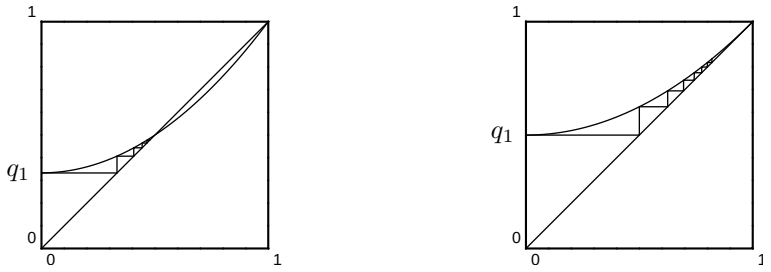

Figure 15.2: Generating functions for case 1 and case 2 respectively.

Case 1. The graph of $f(x)$ first crosses the line $y = x$ at a point $a < 1$. In this case a is the unique fixed point for $f(x)$ in $(0,1)$ and q_n converges to a. (This can be seen by iterating f graphically, as in the left-hand panel in Figure 15.2.) Hence the extinction probability q^* is a. This corresponds to the second case in Theorem 15.3 above.

Case 2. The graph of $f(x)$ first crosses the line $y = x$ at 1. Note that $f(1) = 1$. $f(x)$ does not have a fixed point in $(0,1)$. In this case q_n converges to 1 and the extinction probability q^* is 1. This corresponds to the first case in Theorem 15.3 above.

The two cases above are distinguished by the derivative of $f(x)$ evaluated at $x = 1$. Note that $f'(1) = E[X]$. If $E[X] > 1$ we are in case 1, while if $E[X] < 1$ we are in case 2. The statement of Theorem 15.3 follows.

15.2.3 Galton-Watson in Random Graphs

The number of neighbors of a node v in $G \in \mathcal{G}_{n,p}$ is distributed as $\text{Bin}(n-1,p)$. Now consider exploring the connected component of v by revealing first the neighbors of v, then the (new) neighbors of each of these neighbors, and so on. This is just like a branching process, except that the number of offspring is not uniformly $\text{Bin}(n-1,p)$ at every node: rather, it is $\text{Bin}(n-m,p)$, where m is the number of nodes we have revealed so far. However, as long as m is not too large we might expect that this difference is not significant, so we can use the branching process based on $X \sim \text{Bin}(n,p)$ to analyze the components of G.

What does Theorem 15.3 tell us about this branching process? Note that since $p = \frac{c}{n}$ we have $E[X] = c$. Hence when $c < 1$ we would expect the process to die out a.s., which is in line with our claim that all components are small in this case. When $c > 1$ we would expect the process to continue for a long time with constant probability, which again is in line with our claim that a constant fraction of the nodes are in a giant component of size αn.

Let's see what the value of α should be. To do this, we use the fact that the probability generating function of $\text{Bin}(n, \frac{c}{n})$ converges pointwise to that of $\text{Poisson}(c)$. Hence the extinction probability for our branching process is the same as for that defined by a $\text{Poisson}(c)$ r.v. But the probability generating function for the latter is

$$f(x) \quad = \quad \sum_i \frac{c^i e^{-c}}{i!} x^i = e^{c(x-1)}.$$

Therefore, the extinction probability p^* is the solution to the equation

$$e^{c(x-1)} = x. \tag{15.1}$$

Writing $\alpha = 1 - p^*$ for the probability that a vertex is in the giant component, this equation becomes $e^{-c\alpha} = 1 - \alpha$, exactly as claimed in Theorem 15.2. This completes our intuition for Theorem 15.2. In the next lecture, we will actually prove it.

References

[ABKU94] Y. Azar, A. Broder, A. Karlin and E. Upfal, "Balanced allocations," in *Proceedings of the 26th ACM Symposium on Theory of Computing*, 1994, pp. 593–602.

[GS] G.R. Grimmett and D.R. Stirzaker, *Probability and Random Processes* (2nd ed.), Oxford Univ Press, 1992.

[Mi96] M. Mitzenmacher, *The Power of Two Choices in Randomized Load Balancing*, PhD Thesis, UC Berkeley, 1996.

[Will] D. Williams, *Probability with Martingales*, Cambridge Univ Press, 1991.