— 7 —

The Birthday Paradox, Occupancy and the Coupon Collector Problem

I built on the sand
And it tumbled down,
I built on a rock
And it tumbled down.
Now when I build, I shall begin
With the smoke from the chimney

Leopold Staff, Foundations

7.1. Some needed math

Lemma 7.1.1. *For any positive integer n, we have the following bounds*[7.1].

(i) $1 + x \le e^x$ *and* $1 - x \le e^{-x}$, *for all* x.

(ii) $(1 + 1/n)^n \le e \le (1 + 1/n)^{n+1}$.

(iii) $(1 - 1/n)^n \le \frac{1}{e} \le (1 - 1/n)^{n-1}$.

(iv) $(n/e)^n \le n! \le (n + 1)^{n+1}/e^n$.

(v) *For any* $k \le n$, *we have:* $\left(\frac{n}{k}\right)^k \le \binom{n}{k} \le \left(\frac{ne}{k}\right)^k$.

Proof: **(i)** Let $h(x) = e^x - 1 - x$. Observe that $h'(x) = e^x - 1$, and $h''(x) = e^x > 0$, for all x. That is, $h(x)$ is a convex function. It achieves its minimum at $h'(x) = 0 \implies e^x = 1$, which is true for $x = 0$. For $x = 0$, we have that $h(0) = e^0 - 1 - 0 = 0$. That is, $h(x) \ge 0$ for all x, which implies that $e^x \ge 1 + x$, see Figure 7.1.

[7.1] Nitty-picky: In (iii), assume $n \ge 2$; in (v), k is an integer with $1 \le k \le n$. (Nitpicking is the official english word, but nitty-picky is so much nicer.)

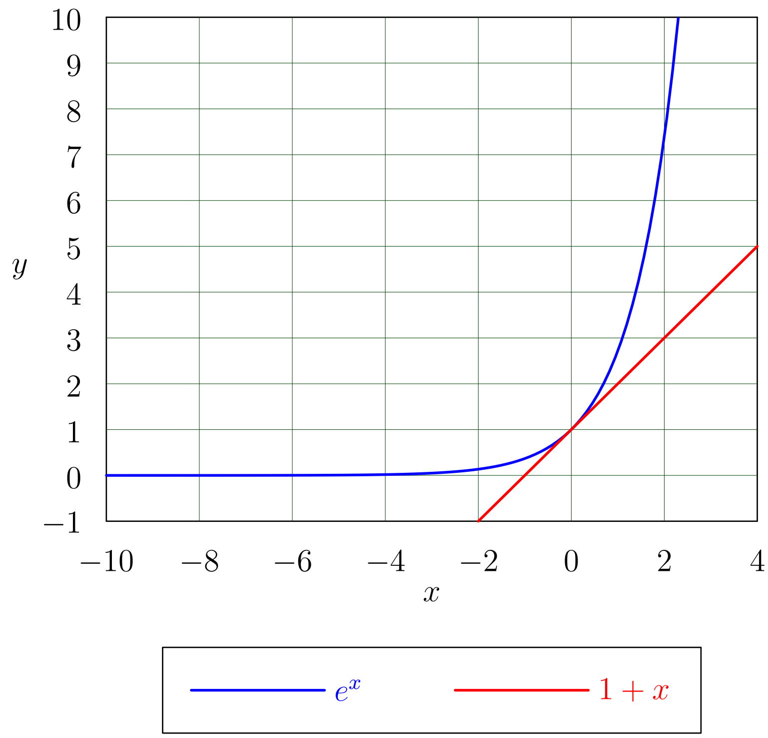

Figure 7.1: The tangent line $y = 1 + x$ lies below the convex curve $y = e^x$.

(ii, iii) Indeed, $1 + 1/n \leq \exp(1/n)$ and $(1 - 1/n) \leq \exp(-1/n)$, by (i). As such

$$\begin{aligned} (1+1/n)^n &\leq \exp(n(1/n)) = e, \\ \text{and} \quad (1-1/n)^n &\leq \exp(n(-1/n)) = \frac{1}{e}, \end{aligned}$$

which implies the left sides of (ii) and (iii). These are equivalent to

$$\frac{1}{e} \leq \left(\frac{n}{n+1}\right)^n = \left(1 - \frac{1}{n+1}\right)^n \quad \text{and} \quad e \leq \left(1 + \frac{1}{n-1}\right)^n,$$

which are the right sides of (iii) [by replacing $n+1$ by n], and the right side of (ii) [by replacing n by $n+1$].

(iv) Indeed,

$$\frac{n^n}{n!} \leq \sum_{i=0}^{\infty} \frac{n^i}{i!} = e^n,$$

by the Taylor expansion of $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$. This implies that $(n/e)^n \leq n!$, as required.

As for the right-hand side, the claim holds for $n = 0$ and $n = 1$. Let $f(n) = (n+1)^{n+1}/e^n$, and observe that by (ii), we have

$$\begin{aligned}\frac{f(n)}{f(n-1)} &= \frac{(n+1)^{n+1}/e^n}{n^n/e^{n-1}} = \frac{n}{e}\left(\frac{n+1}{n}\right)^{n+1} = \frac{n}{e}\left(1+\frac{1}{n}\right)^{n+1}\\ &\geq n\frac{e}{e} = n.\end{aligned}$$

Thus, we have

$$\frac{(n+1)^{n+1}}{e^n} = f(n) = \frac{f(n)}{f(n-1)} \cdot \frac{f(n-1)}{f(n-2)} \cdots \frac{f(1)}{f(0)} \geq n \cdot (n-1) \cdots 1 = n!$$

(v) For $k = 1$, the lower bound is an equality. For the next comparison, assume $2 \leq k \leq n$. For any $k \leq n$ in this range, we have $\frac{n}{k} \leq \frac{n-1}{k-1}$ since $kn - n = n(k-1) \leq k(n-1) = kn - k$. As such, $\frac{n}{k} \leq \frac{n-i}{k-i}$, for $1 \leq i \leq k-1$. As such,

$$\left(\frac{n}{k}\right)^k \leq \frac{n}{k} \cdot \frac{n-1}{k-1} \cdots \frac{n-i}{k-i} \cdots \frac{n-k+1}{1} = \frac{n!}{(n-k)!k!} = \binom{n}{k}.$$

As for the other direction, by (iv), we have

$$\binom{n}{k} \leq \frac{n^k}{k!} \leq \frac{n^k}{(k/e)^k} = \left(\frac{ne}{k}\right)^k.$$

7.2. The birthday paradox

Consider a group of n people, and assume their birthdays are independently and uniformly distributed over the dates in the year (this assumption is not quite true, but close enough). We are interested in the question of how large n has to be till we get a collision – that is, two people with the same birthday. Intuitively, since the year has $m = 365$ days[7.1] , the

[7.1]Leap years have 366 days (every four years, roughly). Moon years have 354 days, which means you have to add a leap month roughly every 19/7 years, which is what the Jewish calendar does. If you do not do this, then the calendar is not synchronized with the weather (this is what the Islamic calendar does). Stick to the Gregorian calendar kids—it is much better.

probability of a person to land on a specific birthday is $p = 1/365$. So the natural guess would be that n needs to be approximately 365. Surprisingly, the answer is much smaller.

Lemma 7.2.1. *Let $X_1, \ldots, X_n$ be n variables picked uniformly, randomly and independently from $[\![m]\!] = \{1, \ldots, m\}$. Then, the expected number of collisions is $\binom{n}{2}/m$, where a* **collision** *is an unordered pair of indices with equal values.*

Proof: Let $Y_{i,j} = 1 \iff X_i = X_j$. We have that $\mathbb{E}[Y_{i,j}] = \mathbb{P}[Y_{i,j} = 1] = 1/m$. Thus, the expected number of collisions is

$$\mathbb{E}\left[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Y_{i,j}\right] = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}[Y_{i,j}] = \binom{n}{2}\frac{1}{m}.$$

Remark 7.2.2. (A) For birthdays, for $m = 365$, and $n = 28$, we have that the expected number of collisions is

$$\binom{28}{2}\frac{1}{365} = \frac{378}{365} > 1.$$

Namely, in expectation, in a room with 28 people we expect at least one pair of people to share a birthday.

(B) If a value is repeated $k > 1$ times, then there are $\binom{k}{2}$ associated collisions.

An expected count above one does not by itself establish a large probability of a collision. The next lemma bounds that probability by bounding its complement.

Lemma 7.2.3. *Let $X_1, \ldots, X_n$ be n variables picked uniformly, randomly and independently from $[\![m]\!] = \{1, \ldots, m\}$. Then, the probability that no collision happened is at most* $\exp(-\binom{n}{2}/m)$.

Proof: If $n > m$, a collision is certain by the pigeonhole principle, and the bound is immediate. Hence assume $1 \leq n \leq m$, and let $\mathcal{B}_0$ be the sure event (i.e., it is an empty event that always happens, and its probability is one). All conditioning events below then have positive probability. Let $\mathcal{E}_i$ be the event that X_i is distinct from all the values in $X_1, \ldots, X_{i-1}$. Let

$$\mathcal{B}_k \quad = \quad \mathcal{E}_1 \cap \mathcal{E}_2 \cap \cdots \cap \mathcal{E}_k \quad = \quad \mathcal{B}_{k-1} \cap \mathcal{E}_k$$

be the event that all of $X_1, \ldots, X_k$ are distinct. Clearly, we have

$$\begin{aligned}
\mathbb{P}[\mathcal{E}_i \mid \mathcal{B}_{i-1}] &= \mathbb{P}[\mathcal{E}_i \mid \mathcal{E}_1 \cap \cdots \cap \mathcal{E}_{i-1}] = \frac{m-(i-1)}{m}\\
&= 1 - \frac{i-1}{m}\\
&\leq \exp\left(-\frac{i-1}{m}\right).
\end{aligned}$$

Observe that

$$\begin{aligned}
\mathbb{P}[\mathcal{B}_i] &= \mathbb{P}\left[\overbrace{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \cdots \cap \mathcal{E}_{i-1}}^{\mathcal{B}_{i-1}} \cap \mathcal{E}_i\right]\\
&= \mathbb{P}[\mathcal{B}_{i-1}] \frac{\mathbb{P}[\mathcal{E}_i \cap \mathcal{B}_{i-1}]}{\mathbb{P}[\mathcal{B}_{i-1}]}\\
&= \mathbb{P}[\mathcal{B}_{i-1}]\, \mathbb{P}[\mathcal{E}_i \mid \mathcal{B}_{i-1}] && \text{(By def of cond. prob)}\\
&\leq \exp\left(-\frac{i-1}{m}\right) \mathbb{P}[\mathcal{B}_{i-1}]\\
&\leq \prod_{k=1}^{i} \exp\left(-\frac{k-1}{m}\right).\\
&= \exp\left(-\sum_{k=1}^{i} \frac{k-1}{m}\right)\\
&= \exp\left(-\frac{i(i-1)}{2}\frac{1}{m}\right)\\
&= \exp\left(-\binom{i}{2}/m\right).
\end{aligned}$$

This implies the desired claim for $i = n$.

7.3. Occupancy problems

Problem 7.3.1. We are throwing m balls into n bins randomly (i.e., for every ball we randomly and uniformly pick a bin independently of all other balls from the n available bins, and place the ball in the bin picked). There are many natural questions one can ask here:

(A) What is the maximum number of balls in any bin?

(B) What is the number of bins which are empty?

(C) How many balls do we have to throw, such that all the bins are non-empty, with reasonable probability?

Example 7.3.2. Consider throwing n balls into n bins, with $n \geq 2$, and consider a specific ball. What is the probability that this ball is all alone in its own bin at the end of the process? Well, we might as well think about throwing this ball first, and all other balls later (by symmetry). Then, all it needs is that all other balls avoid its bin. The probability for that is

$$p = \left(1 - \frac{1}{n}\right)^{n-1} \approx \frac{1}{e}.$$

Thus, if X_i is an indicator variable for the ith ball being lonely, then the expected number of lonely balls (and thus the number of bins with exactly on ball) is

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np \approx \frac{n}{e}.$$

Namely, a constant fraction of the balls are lonely in expectation.

7.3.1. Understanding pileups

The ***load*** of a bin is the number of balls in the bin. Here we are interested in the maximum load—what is the maximum number of balls in a single bin?

7.3.1.1. Intuition

Observation 7.3.3. *An alternative interesting question is how many bins are not empty (in expectation). The probability of a bin to be empty is* $(1 - 1/n)^n \leq 1/e$*, so by linearity of expectation the expected number of empty bins is* $\leq n/e$.

To understand the process better, imagine throwing in the balls. When we throw a ball and it falls into a non-empty bin, we mark it as "reject" and put it into a set of rejected balls R_1. We do this until all balls have been thrown. We now take all the rejected balls and throw them into the bins, but restrict them to bins that contain at least one ball. Again, if a ball falls into a bin that already has two balls, it is rejected into a set R_2. In the ith iteration of this process, all the balls in R_{i-1} are thrown into the bins that have exactly $i-1$ balls in them (no bin has more balls, since such balls were rejected into R_{i-1}). Again, the process rejects all the balls that try to be put into bins that already have exactly i balls, putting them into the set R_i. It is important at this point to convince yourself that this process is indeed equivalent to the original process.

We are going to simplify things a bit—and thus get a loose bound, which is intuitively correct (it is formally correct, but arguing its correctness in detail is not worth the energy at this point). Imagine that we are always throwing the R_{i-1} balls into $Y_{i-1} = |R_{i-1}|$ bins (the real number of bins being used would typically be much bigger). By the above, we have

$$\mathbb{E}[Y_i \mid Y_{i-1}] \;\leq\; Y_{i-1}/e.$$

To see the intuition behind this, observe that the number of bins that still have only $i-1$ balls at the end of this process is, in expectation, at most Y_{i-1}/e, by Observation 7.3.3 (since all the bins under consideration already have $i-1$ balls). Thus, in expectation, at least

$$\Delta_i \;=\; (1-1/e)Y_{i-1}$$

balls are used in creating the ith layer, and hence

$$Y_i \;\leq\; Y_{i-1} - \Delta_i \;=\; Y_{i-1}/e.$$

Important! These calculations are really statements about expectations, and we are being fast and loose with the math to keep the discussion clear. Thus,

$$\mathbb{E}[Y_i] \;=\; \mathbb{E}\Big[\mathbb{E}[Y_i \mid Y_{i-1}]\Big] \;\leq\; \mathbb{E}[Y_{i-1}/e] \;=\; \frac{1}{e}\mathbb{E}[Y_{i-1}]\,\&\;\leq\; \frac{1}{e^i}n.$$

Namely, after $O(\log n)$ iterations, the expected number of rejected balls becomes arbitrarily small. With high probability, after $O(\log n)$ iterations there are no rejected balls left. In other words, all bins have load at most $O(\log n)$.

7.3.1.2. Just doing the math

One can prove a better bound, but that requires doing the exact calculations.

Example 7.3.4 (Heavy times.). A ball is ***k-heavy*** if it there are $k-1$ or more balls that collide with it. What is the expected number of k-heavy balls? For this to happen, there must be $k-1$ balls that "conspire" to collide with this ball. The probability for that is $1/n^{k-1}$ for this specific group of $k-1$ balls, and if so, it creates one k-heavy ball. There are at most $\binom{n-1}{k-1}$ such groups, so by the union bound, the probability of a specific ball to be k-heavy is at most

$$p_{k-1} = \frac{\binom{n-1}{k-1}}{n^{k-1}}.$$

For problem , the above question suggests the bound XXX

Suppose n balls are placed independently and uniformly into n bins, with n sufficiently large. How many ball

Theorem 7.3.5. *Suppose n balls are placed independently and uniformly into n bins, with n sufficiently large. With probability at least $1-1/n$, no bin has more than $k^* = \left\lceil \dfrac{3 \ln n}{\ln \ln n} \right\rceil$ balls in it.*

Proof: Let X_i be the number of balls in the ith bin, when we throw n balls into n bins (i.e., $m = n$). Clearly,

$$\mathbb{E}[X_i] = \sum_{j=1}^{n} \mathbb{P}[\text{The } j\text{th ball falls in the } i\text{th bin}] = n \cdot \frac{1}{n} = 1,$$

by linearity of expectation. The probability that the first bin has exactly i balls is

$$\begin{aligned} p_i &= \binom{n}{i}\left(\frac{1}{n}\right)^i \left(1 - \frac{1}{n}\right)^{n-i} \leq \binom{n}{i}\left(\frac{1}{n}\right)^i \\ &\leq \left(\frac{ne}{i}\right)^i \left(\frac{1}{n}\right)^i = \left(\frac{e}{i}\right)^i \end{aligned}$$

This follows by Lemma 7.1.1 (v).

Let $C(k)$ be the event that the first bin (but really any specific bin, as they all behave in the same way by symmetry) has k or more balls in it. Then,

$$\begin{aligned}\mathbb{P}[C(k)] &\le \sum_{i=k}^{n} p_i \le \sum_{i=k}^{n}\left(\frac{e}{i}\right)^i \le \left(\frac{e}{k}\right)^k\left(1+\frac{e}{k}+\frac{e^2}{k^2}+\ldots\right)\\ &= \left(\frac{e}{k}\right)^k \frac{1}{1-e/k}.\end{aligned}$$

For $k^* = c\ln n/\ln\ln n$, we have

$$\begin{aligned}\mathbb{P}[C(k^*)] &\le \left(\frac{e}{k^*}\right)^{k^*}\frac{1}{1-e/k^*}\\ &\le 2\exp\Big(k^*(1-\ln k^*)\Big) \le 2\exp\left(-\frac{k^*\ln k^*}{2}\right)\\ &\le 2\exp\left(-\frac{c\ln n}{2\ln\ln n}\ln\underbrace{\frac{c\ln n}{\ln\ln n}}_{\approx \ln\ln n}\right)\\ &\le 2\exp\left(-\frac{c\ln n}{4}\right) \le \frac{2}{n^{c/4}}, \le \frac{1}{n^2},\end{aligned}$$

for sufficiently large n and $c \ge 8$.

Let us redo this calculation more carefully (yuk!). For $k^* = \left\lceil \frac{3\ln n}{\ln\ln n}\right\rceil$, we have

$$\begin{aligned}\mathbb{P}[C(k^*)] &\le \left(\frac{e}{k^*}\right)^{k^*}\frac{1}{1-e/k^*}\\ &\le 2\left(\frac{e}{(3\ln n)/\ln\ln n}\right)^{k^*}\\ &= 2\exp\Big(\underbrace{1-\ln 3}_{<0} - \ln\ln n + \ln\ln\ln n\Big)^{k^*}\\ &\le 2\exp((-\ln\ln n + \ln\ln\ln n)k^*)\\ &\le 2\exp\left(-3\ln n + 6\ln n\frac{\ln\ln\ln n}{\ln\ln n}\right)\\ &\le 2\exp(-2.5\ln n)\\ &\le \frac{1}{n^2},\end{aligned}$$

for n large enough. We conclude, by the union bound, that since there are n bins with identical distributions,

$$\mathbb{P}[\text{any bin contains more than } k^* \text{ balls}] \leq \sum_{i=1}^{n} C_i(k^*) \leq \frac{1}{n},$$

where $C_i(k)$ is the event that bin i contains at least k balls.

Exercise 7.3.6. Show that when throwing $m = n \ln n$ balls into n bins, with probability $1 - o(1)$, every bin has $O(\log n)$ balls.

7.3.2. Probability of all bins to have exactly one ball

Next, we are interested in the probability that all m balls fall in distinct bins. Let X_i be the event that the ith ball fell in a distinct bin from the first $i-1$ balls. We have:

$$\begin{aligned}
\mathbb{P}[\cap_{i=2}^{m} X_i] &= \mathbb{P}[X_2] \prod_{i=3}^{m} \mathbb{P}\left[X_i \mid \cap_{j=2}^{i-1} X_j\right] \\
&\leq \prod_{i=2}^{m} \left(\frac{n-i+1}{n}\right) \leq \prod_{i=2}^{m} \left(1 - \frac{i-1}{n}\right) \\
&\leq \prod_{i=2}^{m} e^{-(i-1)/n} \leq \exp\left(-\frac{m(m-1)}{2n}\right),
\end{aligned}$$

thus for $m = \lceil \sqrt{2n} + 1 \rceil$, the probability that all the m balls fall in different bins is smaller than $1/e$.

This is sometimes referred to as the *birthday paradox*, which was already mentioned above. You have $m = 30$ people in the room, and you ask them for the date (day and month) of their birthday (i.e., $n = 365$). The above shows that the probability of all birthdays to be distinct is at most $\exp(-30 \cdot 29/730) \leq 1/e$. Namely, there is more than 50% chance for a birthday collision, a simple but counter-intuitive phenomenon.

7.4. The Coupon Collector's Problem

There are n types of coupons, and at each trial one coupon is picked independently and uniformly at random. How many trials does one have to perform before picking all coupons? Let m be the number of trials performed. We would like to bound the probability that m exceeds a certain number, and we still did not pick all coupons.

Let $C_i \in \{1, \dots, n\}$ be the coupon picked in the ith trial. The jth trial is a success, if C_j was not picked before in the first $j-1$ trials. Let X_i denote the number of trials strictly after the ith success, up to and including the $(i+1)$th success. Let $X_0 = 1$. Clearly, the number of trials performed is

$$X = \sum_{i=0}^{n-1} X_i.$$

Furthermore, X_i has a geometric distribution with parameter p_i, that is $X_i \sim \text{Geom}(p_i)$, with $p_i = (n-i)/n$. The expectation and variance of X_i are

$$\mathbb{E}[X_i] = \frac{1}{p_i} \qquad \text{and} \qquad \mathbb{V}[X_i] = \frac{1-p_i}{p_i^2}.$$

Lemma 7.4.1. *Let X be the number of rounds till we collect all n coupons. Then, $\mathbb{V}[X] \approx (\pi^2/6)n^2$ and its standard deviation is $\sigma_X \approx (\pi/\sqrt{6})n$.*

Proof: The probability of X_i to succeed in a trial is $p_i = (n-i)/n$, and X_i has the geometric distribution with probability p_i. As such $\mathbb{E}[X_i] = 1/p_i$, and $\mathbb{V}[X_i] = (1-p_i)/p_i^2$. Thus,

$$\mathbb{E}[X] = \sum_{i=0}^{n-1} \mathbb{E}[X_i] = \sum_{i=0}^{n-1} \frac{n}{n-i} = nH_n = n(\ln n + \Theta(1)) = n \ln n + O(n),$$

where $H_n = \sum_{i=1}^{n} 1/i$ is the nth harmonic number.

As for variance, using the independence of $X_0, \dots, X_{n-1}$, we have

$$\begin{aligned}
\mathbb{V}[X] &= \sum_{i=0}^{n-1} \mathbb{V}[X_i] \\
&= \sum_{i=0}^{n-1} \frac{1-p_i}{p_i^2} \\
&= \sum_{i=0}^{n-1} \frac{1-(n-i)/n}{\left(\frac{n-i}{n}\right)^2} \quad = \quad \sum_{i=0}^{n-1} \frac{i/n}{\left(\frac{n-i}{n}\right)^2}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=0}^{n-1} \frac{i}{n}\left(\frac{n}{n-i}\right)^2 \\
&= n\sum_{i=0}^{n-1} \frac{i}{(n-i)^2} = n\sum_{i=1}^{n} \frac{n-i}{i^2} \\
&= n\left(\sum_{i=1}^{n} \frac{n}{i^2} - \sum_{i=1}^{n} \frac{1}{i}\right) = n^2 \sum_{i=1}^{n} \frac{1}{i^2} - nH_n \approx \frac{\pi^2}{6} n^2,
\end{aligned}
$$

since $\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{i^2} = \pi^2/6$, we have $\displaystyle\lim_{n\to\infty} \frac{\mathbb{V}[X]}{n^2} = \frac{\pi^2}{6}$.

This implies a weak bound on the concentration of X, using Chebyshev's inequality. We have

$$
\mathbb{P}\left[X \geq n \ln n + n + t \cdot n\frac{\pi}{\sqrt{6}}\right] \leq \mathbb{P}\left[\left|X - \mathbb{E}[X]\right| \geq t\sigma_X\right] \leq \frac{1}{t^2},
$$

7.5. Notes

The material in this note covers parts of [MR95, sections 3.1, 3.2, 3.6].

References

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge, UK: Cambridge University Press, 1995.