21 Random Walks

Random Walks are used to model situations in which an object moves in a sequence of steps in randomly chosen directions. For example, physicists use three-dimensional random walks to model Brownian motion and gas diffusion. In this chapter we'll examine two examples of random walks. First, we'll model gambling as a simple 1-dimensional random walk—a walk along a straight line. Then we'll explain how the Google search engine used random walks through the graph of world-wide web links to determine the relative importance of websites.

21.1 Gambler's Ruin

Suppose a gambler starts with an initial stake of n dollars and makes a sequence of \$1 bets. If he wins an individual bet, he gets his money back plus another \$1. If he loses the bet, he loses the \$1.

We can model this scenario as a random walk between integer points on the real line. The position on the line at any time corresponds to the gambler's cash-on-hand, or *capital*. Walking one step to the right corresponds to winning a \$1 bet and thereby increasing his capital by \$1. Similarly, walking one step to the left corresponds to losing a \$1 bet.

The gambler plays until either he runs out of money or increases his capital to a target amount of T dollars. The amount $T - n$ is defined to be his *intended profit*.

If he reaches his target, he will have won his intended profit and is called an overall *winner*. If his capital reaches zero before reaching his target, he will have lost n dollars; this is called *going broke* or being *ruined*. We'll assume that the gambler has the same probability p of winning each individual \$1 bet, and that the bets are mutually independent. We'd like to find the probability that the gambler wins.

The gambler's situation as he proceeds with his \$1 bets is illustrated in Figure 21.1. The random walk has boundaries at 0 and T. If the random walk ever reaches either of these boundary values, then it terminates.

In an *unbiased game*, the individual bets are fair: the gambler is equally likely to win or lose each bet—that is, $p = 1/2$. The gambler is more likely to win if $p > 1/2$ and less likely to win if $p < 1/2$; these random walks are called *biased*. We want to determine the probability that the walk terminates at boundary T—the probability that the gambler wins. We'll do this in Section 21.1.1. But before we

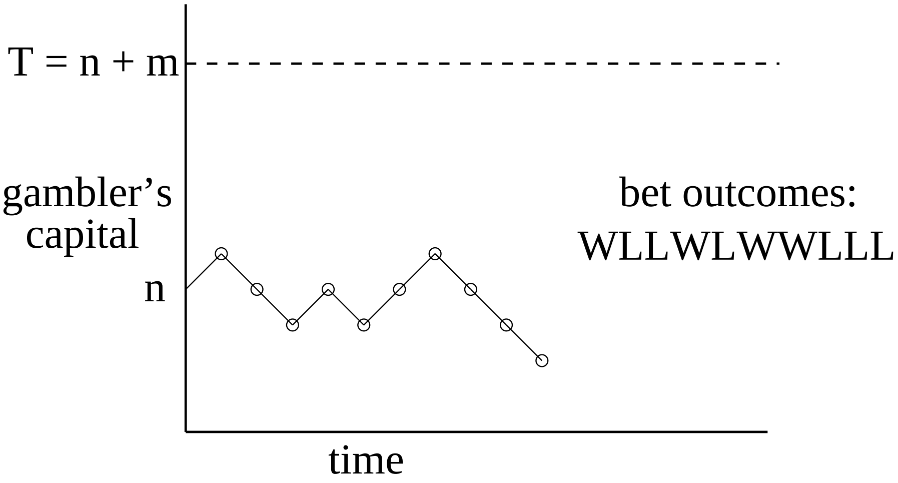


Figure 21.1 A graph of the gambler’s capital versus time for one possible sequence of bet outcomes. At each time step, the graph goes up with probability p and down with probability $1 - p$. The gambler continues betting until the graph reaches either 0 or T. If he starts with $\$n$, his intended profit is $\$m$ where $T = n + m$.

derive the probability, let’s examine what it turns out to be.

Let’s begin by supposing the gambler plays an unbiased game starting with $100 and will play until he goes broke or reaches a target of 200 dollars. Since he starts equidistant from his target and bankruptcy in this case, it’s clear by symmetry that his probability of winning is 1/2.

We’ll show below that starting with n dollars and aiming for a target of $T \geq n$ dollars, the probability the gambler reaches his target before going broke is n/T. For example, suppose he wants to win the same $100, but instead starts out with $500. Now his chances are pretty good: the probability of his making the 100 dollars is $5/6$. And if he started with one million dollars still aiming to win $100 dollars he almost certain to win: the probability is $1M/(1M + 100) > .9999$.

So in the unbiased game, the larger the initial stake relative to the target, the higher the probability the gambler will win, which makes some intuitive sense. But note that although the gambler now wins nearly all the time, when he loses, he loses *big*. Bankruptcy costs him $1M, while when he wins, he wins only $100. The gambler’s average win remains zero dollars, which is what you’d expect when making fair bets.

Another useful way to describe this scenario is as a game between two players. Say Albert starts with $500, and Eric starts with $100. They flip a fair coin, and every time a Head appears, Albert wins \$1 from Eric, and vice versa for Tails. They play this game until one person goes bankrupt. This problem is identical to the Gambler's Ruin problem with $n = 500$ and $T = 100 + 500 = 600$. The probability of Albert winning is $500/600 = 5/6$.

Now suppose instead that the gambler chooses to play roulette in an American casino, always betting \$1 on red. Because the casino puts two green numbers on its roulette wheels, the probability of winning a single bet is a little less than 1/2. The casino has an advantage, but the bets are close to fair, and you might expect that starting with \$500, the gambler has a reasonable chance of winning \$100—the 5/6 probability of winning in the unbiased game surely gets reduced, but perhaps not too drastically.

This mistaken intuition is how casinos stay in business. In fact, the gambler's odds of winning \$100 by making \$1 bets against the "slightly" unfair roulette wheel are less than 1 in 37,000. If that's surprising to you, it only gets weirder from here: 1 in 37,000 is in fact an upper bound on the gambler's chance of winning *regardless of his starting stake*. Whether he starts with \$5000 or \$5 billion, he still has almost no chance of winning!

21.1.1 The Probability of Avoiding Ruin

We will determine the probability that the gambler wins using an idea of Pascal's dating back to the beginnings probability theory in the mid-seventeenth century.

Pascal viewed the walk as a two-player game between Albert and Eric as described above. Albert starts with a stack of n chips and Eric starts with a stack of $m = T - n$ chips. At each bet, Albert wins Eric's top chip with probability p and loses his top chip to Eric with probability $q ::= 1 - p$. They play this game until one person goes bankrupt.

Pascal's ingenious idea was to alter the worth of the chips to make the game fair regardless of p. Specifically, Pascal assigned Albert's bottom chip a worth of $r ::= q/p$ and then assigned successive chips *up* his stack worths equal to $r^2, r^3, \ldots$ up to his top chip with worth r^n. Eric's top chip gets assigned worth r^{n+1}, and the successive chips *down* his stack are worth $r^{n+2}, r^{n+3}, \ldots$ down to his bottom chip worth r^{n+m}.

The expected payoff of Albert's first bet is worth

$$r^{n+1} \cdot p - r^n \cdot q = \left(r^n \cdot \frac{q}{p}\right) \cdot p - r^n \cdot q = 0.$$

so this assignment makes the first bet a fair one in terms of worth. Moreover, whether Albert wins or loses the bet, the successive chip worths counting up Albert's stack and then down Eric's remain $r, r^2, \ldots, r^n, \ldots, r^{n+m}$, ensuring by the same reasoning that every bet has fair worth. So, Albert’s expected worth at the end of the game is the sum of the expectations of the worth of each bet, which is 0.[1]

When Albert wins all of Eric’s chips his total gain is worth

$$\sum_{i=n+1}^{n+m} r^i,$$

and when he loses all his chips to Eric, his total loss is worth $\sum_{i=1}^{n} r^i$. Letting w_n be Albert’s probability of winning, we now have

$$0 = \text{Ex[worth of Albert's payoff]} = w_n \sum_{i=n+1}^{n+m} r^i - (1 - w_n) \sum_{i=1}^{n} r^i.$$

In the truly fair game when $r = 1$, we have $0 = mw_n - n(1 - w_n)$, so $w_n = n/(n + m)$, as claimed above.

In the biased game with $r \neq 1$, we have

$$0 = r \cdot \frac{r^{n+m} - r^n}{r - 1} \cdot w_n - r \cdot \frac{r^n - 1}{r - 1} \cdot (1 - w_n).$$

Solving for w_n gives

$$w_n = \frac{r^n - 1}{r^{n+m} - 1} = \frac{r^n - 1}{r^T - 1} \tag{21.1}$$

We have now proved

Theorem 21.1.1. *In the Gambler’s Ruin game with initial capital n, target T, and probability p of winning each individual bet,*

$$\Pr[\textit{the gambler wins}] = \begin{cases} \dfrac{n}{T} & \textit{for } p = \dfrac{1}{2}, \\[2ex] \dfrac{r^n - 1}{r^T - 1} & \textit{for } p \neq \dfrac{1}{2}, \end{cases} \tag{21.2}$$

where $r ::= q/p$.

[1] Here we’re legitimately appealing to infinite linearity, since the payoff amounts remain bounded independent of the number of bets.

21.1.2 A Recurrence for the Probability of Winning

Fortunately, you don't need to be as ingenuious Pascal in order to handle Gambler's Ruin, because linear recurrences offer a methodical approach to the basic problems.

The probability that the gambler wins is a function of his initial capital n his target $T \geq n$ and the probability p that the wins an individual one dollar bet. For fixed p and T, let w_n be the gambler's probability of winning when his initial capital is n dollars. For example, w_0 is the probability that the gambler will win given that he starts off broke and w_T is the probability he will win if he starts off with his target amount, so clearly

$$w_0 = 0, \tag{21.3}$$

$$w_T = 1. \tag{21.4}$$

Otherwise, the gambler starts with n dollars, where $0 < n < T$. Now suppose the gambler wins his first bet. In this case, he is left with $n + 1$ dollars and becomes a winner with probability w_{n+1}. On the other hand, if he loses the first bet, he is left with $n - 1$ dollars and becomes a winner with probability w_{n-1}. By the Total Probability Rule, he wins with probability $w_n = pw_{n+1} + qw_{n-1}$. Solving for w_{n+1} we have

$$w_{n+1} = \frac{w_n}{p} - rw_{n-1} \tag{21.5}$$

where r is q/p as in Section 21.1.1.

This recurrence holds only for $n + 1 \leq T$, but there's no harm in using (21.5) to define w_{n+1} for all $n + 1 > 1$. Now, letting

$$W(x) ::= w_0 + w_1 x + w_2 x^2 + \cdots$$

be the generating function for the w_n, we derive from (21.5) and (21.3) using our generating function methods that

$$W(x) = \frac{w_1 x}{rx^2 - x/p + 1}. \tag{21.6}$$

But it's easy to check that the denominator factors:

$$rx^2 - \frac{x}{p} + 1 = (1 - x)(1 - rx).$$

Now if $p \neq q$, then using partial fractions we conclude that

$$W(x) = \frac{A}{1 - x} + \frac{B}{1 - rx}, \tag{21.7}$$

for some constants A, B. To solve for A, B, note that by (21.6) and (21.7),

$$w_1 x = A(1 - rx) + B(1 - x),$$

so letting $x = 1$, we get $A = w_1/(1 - r)$, and letting $x = 1/r$, we get $B = w_1/(r - 1)$. Therefore,

$$W(x) = \frac{w_1}{r - 1}\left(\frac{1}{1 - rx} - \frac{1}{1 - x}\right),$$

which implies

$$w_n = w_1 \frac{r^n - 1}{r - 1}. \tag{21.8}$$

Finally, we can use (21.8) to solve for w_1 by letting $n = T$ to get

$$w_1 = \frac{r - 1}{r^T - 1}.$$

Plugging this value of w_1 into (21.8), we arrive at the solution:

$$w_n = \frac{r^n - 1}{r^T - 1},$$

matching Pascal's result (21.1).

In the unbiased case where $p = q$, we get from (21.6) that

$$W(x) = \frac{w_1 x}{(1 - x)^2},$$

and again can use partial fractions to match Pascal's result (21.2).

21.1.3 A simpler expression for the biased case

The expression (21.1) for the probability that the Gambler wins in the biased game is a little hard to interpret. There is a simpler upper bound which is nearly tight when the gambler's starting capital is large and the game is biased *against* the gambler. Then $r > 1$, both the numerator and denominator in (21.1) are positive, and the numerator is smaller. This implies that

$$w_n < \frac{r^n}{r^T} = \left(\frac{1}{r}\right)^{T-n}$$

and gives:

Corollary 21.1.2. *In the Gambler’s Ruin game with initial capital n, target T, and probability $p < 1/2$ of winning each individual bet,*

$$\Pr[\textit{the gambler wins}] < \left(\frac{1}{r}\right)^{T-n} \tag{21.9}$$

where $r ::= q/p > 1$.

So the gambler gains his intended profit before going broke with probability at most $1/r$ raised to the intended profit power. Notice that this upper bound does not depend on the gambler’s starting capital, but only on his intended profit. This has the amazing consequence we announced above: *no matter how much money he starts with*, if he makes \$1 bets on red in roulette aiming to win \$100, the probability that he wins is less than

$$\left(\frac{18/38}{20/38}\right)^{100} = \left(\frac{9}{10}\right)^{100} < \frac{1}{37,648}.$$

The bound (21.9) decreases exponentially as the intended profit increases. So, for example, doubling his intended profit will square his probability of winning. In this case, the probability that the gambler’s stake goes up 200 dollars before he goes broke playing roulette is at most

$$(9/10)^{200} = ((9/10)^{100})^2 < \left(\frac{1}{37,648}\right)^2,$$

which is about 1 in 1.4 billion.

Intuition

Why is the gambler so unlikely to make money when the game is only slightly biased against him? To answer this intuitively, we can identify two forces at work on the gambler’s wallet. First, the gambler’s capital has random upward and downward *swings* from runs of good and bad luck. Second, the gambler’s capital will have a steady, downward *drift*, because the negative bias means an average loss of a few cents on each \$1 bet. The situation is shown in Figure 21.2.

Our intuition is that if the gambler starts with, say, a billion dollars, then he is sure to play for a very long time, so at some point there should be a lucky, upward swing that puts him \$100 ahead. But his capital is steadily drifting downward. If the gambler does not have a lucky, upward swing early on, then he is doomed. After his capital drifts downward by tens and then hundreds of dollars, the size of the upward swing the gambler needs to win grows larger and larger. And as the

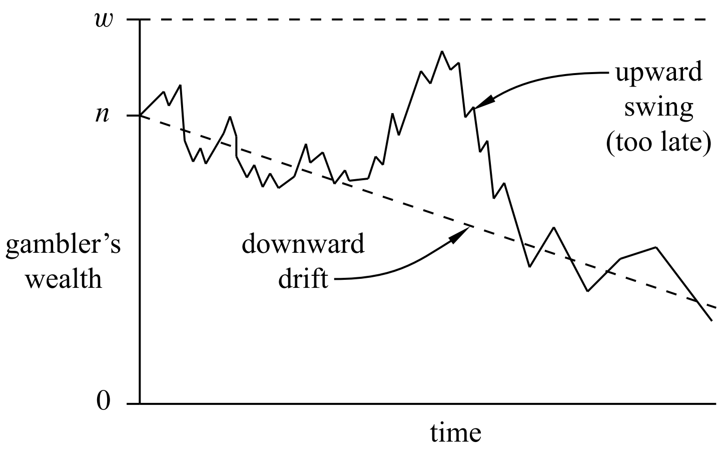


Figure 21.2 In a biased random walk, the downward drift usually dominates swings of good luck.

size of the required swing grows, the odds that it occurs decrease exponentially. As a rule of thumb, *drift dominates swings* in the long term.

We can quantify these drifts and swings. After k rounds for $k \le \min(m, n)$, the number of wins by our player has a binomial distribution with parameters $p < 1/2$ and k. His expected win on any single bet is $p - q = 2p - 1$ dollars, so his expected capital is $n - k(1 - 2p)$. Now to be a winner, his actual number of wins must exceed the expected number by $m + k(1 - 2p)$. But from the formula (20.14), the binomial distribution has a standard deviation of only $\sqrt{kp(1-p)}$. So for the gambler to win, he needs his number of wins to deviate by

$$\frac{m + k(1-2p)}{\sqrt{kp(1-2p)}} = \Theta(\sqrt{k})$$

times its standard deviation. In our study of binomial tails, we saw that this was extremely unlikely.

In a fair game, there is no drift; swings are the only effect. In the absence of downward drift, our earlier intuition is correct. If the gambler starts with a trillion dollars then almost certainly there will eventually be a lucky swing that puts him $100 ahead.

21.1.4 How Long a Walk?

Now that we know the probability w_n that the gambler is a winner in both fair and unfair games, we consider how many bets he needs on average to either win or go broke. A linear recurrence approach works here as well.

For fixed p and T, let e_n be the expected number of bets until the game ends when the gambler's initial capital is n dollars. Since the game is over in zero steps if $n = 0$ or T, the boundary conditions this time are $e_0 = e_T = 0$.

Otherwise, the gambler starts with n dollars, where $0 < n < T$. Now by the conditional expectation rule, the expected number of steps can be broken down into the expected number of steps given the outcome of the first bet weighted by the probability of that outcome. But after the gambler wins the first bet, his capital is $n + 1$, so he can expect to make another e_{n+1} bets. That is,

$$\text{Ex}[\#\text{bets starting with } \$n \mid \text{gambler wins first bet}] = 1 + e_{n+1}.$$

Similarly, after the gambler loses his first bet, he can expect to make another e_{n-1} bets:

$$\text{Ex}[\#\text{bets starting with } \$n \mid \text{gambler loses first bet}] = 1 + e_{n-1}.$$

So we have

$$\begin{aligned} e_n &= p\,\text{Ex}[\#\text{bets starting with } \$n \mid \text{gambler wins first bet}] \\ &\quad + q\,\text{Ex}[\#\text{bets starting with } \$n \mid \text{gambler loses first bet}] \\ &= p(1 + e_{n+1}) + q(1 + e_{n-1}) = pe_{n+1} + qe_{n-1} + 1. \end{aligned}$$

This yields the linear recurrence

$$e_{n+1} = \frac{1}{p}e_n - \frac{q}{p}e_{n-1} - \frac{1}{p}. \tag{21.10}$$

The routine solution of this linear recurrence yields:

Theorem 21.1.3. *In the Gambler's Ruin game with initial capital n, target T, and probability p of winning each bet,*

$$\text{Ex}[\textit{number of bets}] = \begin{cases} n(T-n) & \textit{for } p = \dfrac{1}{2}, \\[2ex] \dfrac{w_n \cdot T - n}{p - q} & \textit{for } p \neq \dfrac{1}{2} \\ & \textit{where } w_n = (r^n - 1)/(r^T - 1) \\ & \qquad = \Pr[\textit{the gambler wins}]. \end{cases} \tag{21.11}$$

In the unbiased case, (21.11) can be rephrased simply as

$$\text{Ex}[\text{number of fair bets}] = \text{initial capital} \cdot \text{intended profit}. \tag{21.12}$$

For example, if the gambler starts with \$10 dollars and plays until he is broke or ahead \$10, then $10 \cdot 10 = 100$ bets are required on average. If he starts with \$500 and plays until he is broke or ahead \$100, then the expected number of bets until the game is over is $500 \times 100 = 50{,}000$. This simple formula (21.12) cries out for an intuitive proof, but we have not found one (where are you, Pascal?).

21.1.5 Quit While You Are Ahead

Suppose that the gambler never quits while he is ahead. That is, he starts with $n > 0$ dollars, ignores any target T, but plays until he is flat broke. Call this the *unbounded Gambler's ruin* game. It turns out that if the game is not favorable, that is, $p \le 1/2$, the gambler is sure to go broke. In particular, this holds in an unbiased game with $p = 1/2$.

Lemma 21.1.4. *If the gambler starts with one or more dollars and plays a fair unbounded game, then he will go broke with probability 1.*

Proof. If the gambler has initial capital n and goes broke in a game without reaching a target T, then he would also go broke if he were playing and ignored the target. So the probability that he will lose if he keeps playing without stopping at any target T must be at least as large as the probability that he loses when he has a target $T > n$.

But we know that in a fair game, the probability that he loses is $1 - n/T$. This number can be made arbitrarily close to 1 by choosing a sufficiently large value of T. Hence, the probability of his losing while playing without any target has a lower bound arbitrarily close to 1, which means it must in fact be 1. ■

So even if the gambler starts with a million dollars and plays a perfectly fair game, he will eventually lose it all with probability 1. But there is good news: if the game is fair, he can "expect" to play forever:

Lemma 21.1.5. *If the gambler starts with one or more dollars and plays a fair unbounded game, then his expected number of plays is infinite.*

A proof appears in Problem 21.2.

So even starting with just one dollar, the expected number of plays before going broke is infinite! This sounds reassuring—you can go about your business without worrying about being doomed, because doom will be infinitely delayed. To illustrate a situation where you really needn't worry, think about mean time to failure with a really tiny probability of failure in any given second—say 10^{-100}. In this case you are unlikely to fail any time much sooner than many lifetimes of the estimated age of the universe, even though you will eventually fail with probability one.

But in general, you shouldn't feel reassured by an infinite expected time to go broke. For example, think about a variant Gambler's Ruin game which works as follows: run one second of the process that has a 10^{-100} of failing in any second. If it does *not* fail, then you go broke immediately. Otherwise, you play a fair, unbounded Gambler's Ruin game. Now there is an overwhelming probability, namely, $1 - 10^{-100}$, that you will go broke immediately. But there is a 10^{-100} probability that you will wind up playing fair Gambler's Ruin, so your overall expected time will be at least 10^{-100} times the expectation of fair Gambler's Ruin, namely, it will still be infinite.

For the actual fair, unbounded Gambler's Ruin gain starting with one dollar, there is a a 50% chance the Gambler will go broke after the first bet, and a more than $15/16$ chance of going broke within five bets, for example. So infinite expected time is not much consolation to a Gambler who goes broke quickly with high probability.

21.2 Random Walks on Graphs

The hyperlink structure of the World Wide Web can be described as a digraph. The vertices are the web pages with a directed edge from vertex x to vertex y if x has a link to y. A digraph showing part of the website for MIT subject 6.042, *Mathematics for Computer Science*, is shown in Figure 21.3.

The web graph is an enormous graph with trillions of vertices. In 1995, two students at Stanford, Larry Page and Sergey Brin, realized that the structure of this graph could be very useful in building a search engine. Traditional document searching programs had been around for a long time and they worked in a fairly straightforward way. Basically, you would enter some search terms and the searching program would return all documents containing those terms. A relevance score might also be returned for each document based on the frequency or position that the search terms appeared in the document. For example, if the search term appeared in the title or appeared 100 times in a document, that document would get a higher score.

This approach works fine if you only have a few documents that match a search term. But on the web, there are many billions of documents and millions of matches to a typical search. For example, on May 2, 2012, a search on Google for " 'Mathematics for Computer Science' text" gave 482,000 hits! Which ones should we look at first? Just because a page gets a high keyword score—say because it has "Mathematics Mathematics ... Mathematics" copied 200 times across the front of the document—does not make it a great candidate for attention. The web is filled with

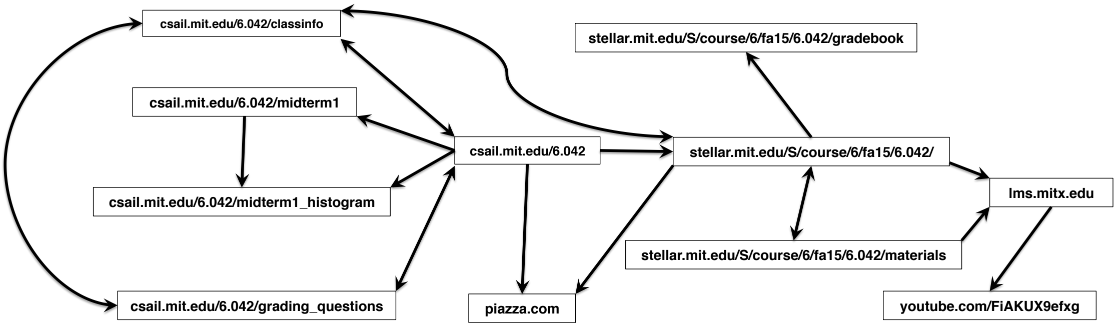

csail.mit.edu/6.042/classinfo
stellar.mit.edu/S/course/6/fa15/6.042/gradebook
csail.mit.edu/6.042/midterm1
csail.mit.edu/6.042
stellar.mit.edu/S/course/6/fa15/6.042/
lms.mitx.edu
csail.mit.edu/6.042/midterm1_histogram
stellar.mit.edu/S/course/6/fa15/6.042/materials
csail.mit.edu/6.042/grading_questions
piazza.com
youtube.com/FiAKUX9efxg

Figure 21.3 Website digraph for MIT subject 6.042

bogus websites that repeat certain words over and over in order to attract visitors.

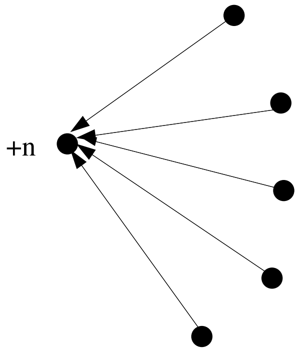
Google’s enormous market capital in part derives from the revenue it receives from advertisers paying to appear at the top of search results. That top placement would not be worth much if Google’s results were as easy to manipulate as keyword frquencies. Advertisers pay because Google’s ranking method is consistently good at determining the most relevant web pages. For example, Google demonstrated its accuracy in our case by giving first rank[2] to our 6.042 text.

So how did Google know to pick our text to be first out of 482,000?—because back in 1995 Larry and Sergey got the idea to allow the digraph structure of the web to determine which pages are likely to be the most important.

21.2.1 A First Crack at Page Rank

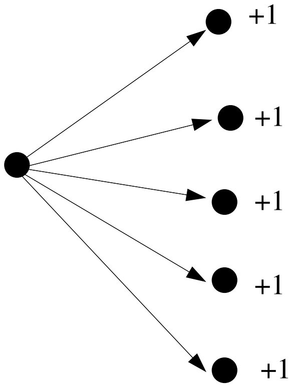
Looking at the web graph, do you have an idea which vertex/page might be the best to rank first? Assume that all the pages match the search terms for now. Well, intuitively, we should choose x_2, since lots of other pages point to it. This leads us to their first idea: try defining the *page rank* of x to be $\text{indegree}(x)$, the number of links pointing to x. The idea is to think of web pages as voting for the most important page—the more votes, the better the rank.

Unfortunately, there are some problems with this idea. Suppose you wanted to have your page get a high ranking. One thing you could do is to create lots of dummy pages with links to your page.



There is another problem—a page could become unfairly influential by having lots of links to other pages it wanted to hype.

[2] First rank for some reason was an early version archived at Princeton; the Spring 2010 version on the MIT Open Courseware site ranked 4th and 5th.



So this strategy for high ranking would amount to, “vote early, vote often,” which is no good if you want to build a search engine that’s worth paying fees for. So, admittedly, their original idea was not so great. It was better than nothing, but certainly not worth billions of dollars.

21.2.2 Random Walk on the Web Graph

But then Sergey and Larry thought some more and came up with a couple of improvements. Instead of just counting the indegree of a vertex, they considered the probability of being at each page after a long random walk on the web graph. In particular, they decided to model a user’s web experience as following each link on a page with uniform probability. For example, if the user is at page x, and there are three links from page x, then each link is followed with probability $1/3$. More generally, they assigned each edge $x \to y$ of the web graph with a probability conditioned on being on page x:

$$\Pr\left[\text{follow link } \langle x \to y \rangle \mid \text{at page } x\right] ::= \frac{1}{\text{outdeg}(x)}.$$

The simulated user experience is then just a random walk on the web graph.

We can also compute the probability of arriving at a particular page y by summing over all edges pointing to y. We thus have

$$\begin{aligned}\Pr[\text{go to } y] &= \sum_{\text{edges } \langle x \to y \rangle} \Pr\left[\text{follow link } \langle x \to y \rangle \mid \text{at page } x\right] \cdot \Pr[\text{at page } x] \\ &= \sum_{\text{edges } \langle x \to y \rangle} \frac{\Pr[\text{at } x]}{\text{outdeg}(x)} \qquad (21.13)\end{aligned}$$

For example, in our web graph, we have

$$\Pr[\text{go to } x_4] = \frac{\Pr[\text{at } x_7]}{2} + \frac{\Pr[\text{at } x_2]}{1}.$$

One can think of this equation as x_7 sending half its probability to x_2 and the other half to x_4. The page x_2 sends all of its probability to x_4.

There's one aspect of the web graph described thus far that doesn't mesh with the user experience—some pages have no hyperlinks out. Under the current model, the user cannot escape these pages. In reality, however, the user doesn't fall off the end of the web into a void of nothingness. Instead, he restarts his web journey. Moreover, even if a user does not get stuck at a dead end, they will commonly get discouraged after following some unproductive path for a while and will decide to restart.

To model this aspect of the web, Sergey and Larry added a *supervertex* to the web graph and added an edge from every page to the supervertex. Moreover, the supervertex points to every other vertex in the graph with equal probability, allowing the walk to restart from a random place. This ensures that the graph is strongly connected.

If a page had no hyperlinks, then its edge to the supervertex has to be assigned probability one. For pages that had some hyperlinks, the additional edge pointing to the supervertex was assigned some specially given probability. In the original versions of Page Rank, this probability was arbitrarily set to 0.15. That is, each vertex with outdegree $n \geq 1$ got an additional edge pointing to the supervertex with assigned probability 0.15; its other n outgoing edges were still kept equally likely, that is, each of the n edges was assigned probability $0.85/n$.

21.2.3 Stationary Distribution & Page Rank

The basic idea behind page rank is finding a stationary distribution over the web graph, so let's define a stationary distribution.

Suppose each vertex is assigned a probability that corresponds, intuitively, to the likelihood that a random walker is at that vertex at a randomly chosen time. We assume that the walk never leaves the vertices in the graph, so we require that

$$\sum_{\text{vertices } x} \Pr[\text{at } x] = 1. \tag{21.14}$$

Definition 21.2.1. An assignment of probabilities to vertices in a digraph is a *stationary distribution* if for all vertices x

$$\Pr[\text{at } x] = \Pr[\text{go to } x \text{ at next step}]$$

Sergey and Larry defined their page ranks to be a stationary distribution. They did this by solving the following system of linear equations: find a nonnegative number Rank(x) for each vertex x such that

$$\text{Rank}(x) = \sum_{\text{edges } \langle y \to x \rangle} \frac{\text{Rank}(y)}{\text{outdeg}(y)}, \tag{21.15}$$

corresponding to the intuitive equations given in (21.13). These numbers must also satisfy the additional constraint corresponding to (21.14):

$$\sum_{\text{vertices } x} \text{Rank}(x) = 1. \tag{21.16}$$

So if there are n vertices, then equations (21.15) and (21.16) provide a system of $n + 1$ linear equations in the n variables Rank(x). Note that constraint (21.16) is needed because the remaining constraints (21.15) could be satisfied by letting Rank$(x) ::= 0$ for all x, which is useless.

Sergey and Larry were smart fellows, and they set up their page rank algorithm so it would always have a meaningful solution. Strongly connected graphs have *unique* stationary distributions (Problem 21.12), and their addition of a supervertex ensures this. Moreover, starting from *any* vertex and taking a sufficiently long random walk on the graph, the probability of being at each page will get closer and closer to the stationary distribution. Note that general digraphs without supervertices may have neither of these properties: there may not be a unique stationary distribution, and even when there is, there may be starting points from which the probabilities of positions during a random walk do not converge to the stationary distribution (Problem 21.8).

Now just keeping track of the digraph whose vertices are trillions of web pages is a daunting task. That's why in 2011 Google invested \$168,000,000 in a solar power plant—the electrical power drawn by Google's servers in 2011 would have supplied the needs of 200,000 households.[3] Indeed, Larry and Sergey named their system Google after the number 10^{100}—which is called a "googol"—to reflect the fact that the web graph is so enormous.

Anyway, now you can see how this text ranked first out of 378,000 matches. Lots of other universities used our notes and presumably have links to the MIT Mathematics for Computer Science Open Course Ware site, and the university sites themselves are legitimate, which ultimately leads to the text getting a high page rank in the web graph.

[3] *Google Details, and Defends, Its Use of Electricity*, New York Times, September 8, 2011.

Problems for Section 21.1

Practice Problems

Problem 21.1.
Suppose that a gambler is playing a game in which he makes a series of \$1 bets. He wins each one with probability 0.49, and he keeps betting until he either runs out of money or reaches some fixed goal of T dollars.

Let $t(n)$ be the expected number of *bets* the gambler makes until the game ends, where n is the number of dollars the gambler has when he starts betting. Then the function t satisfies a linear recurrence of the form

$$t(n) = a \cdot t(n+1) + b \cdot t(n-1) + c$$

for real constants a, b, c, and $0 < n < T$.

(a) What are the values of a, b and c?

(b) What is $t(0)$?

(c) What is $t(T)$?

Class Problems

Problem 21.2.
In a gambler's ruin scenario, the gambler makes independent \$1 bets, where the probability of winning a bet is p and of losing is $q ::= 1 - p$. The gambler keeps betting until he goes broke or reaches a target of T dollars.

Suppose $T = \infty$, that is, the gambler keeps playing until he goes broke. Let r be the probability that starting with $n > 0$ dollars, the gambler's stake ever gets reduced to $n - 1$ dollars.

(a) Explain why

$$r = q + pr^2.$$

(b) Conclude that if $p \leq 1/2$, then $r = 1$.

(c) Prove that even in a fair game, the gambler is sure to get ruined *no matter how much money he starts with*!

(d) Let t be the expected time for the gambler's stake to go down by 1 dollar. Verify that

$$t = q + p(1 + 2t).$$

Conclude that starting with a 1 dollar stake in a fair game, the gambler can expect to play forever!

Problem 21.3.
A gambler is placing \$1 bets on the "1st dozen" in roulette. This bet wins when a number from one to twelve comes in, and then the gambler gets his \$1 back plus \$2 more. Recall that there are 38 numbers on the roulette wheel.

The gambler's initial stake in \$$n$ and his target is \$$T$. He will keep betting until he runs out of money ("goes broke") or reaches his target. Let w_n be the probability of the gambler winning, that is, reaching target \$$T$ before going broke.

(a) Write a linear recurrence with boundary conditions for w_n. You need *not* solve the recurrence.

(b) Let e_n be the expected number of bets until the game ends. Write a linear recurrence with boundary conditions for e_n. You need *not* solve the recurrence.

Problem 21.4.
In the fair Gambler's Ruin game with initial stake of n dollars and target of T dollars, let e_n be the number of \$1 bets the gambler makes until the game ends (because he reaches his target or goes broke).

(a) Describe constants a, b, c such that

$$e_n = ae_{n-1} + be_{n-2} + c. \qquad (21.17)$$

for $1 < n < T$.

(b) Let e_n be defined by (21.17) for all $n > 1$, where $e_0 = 0$ and $e_1 = d$ for some constant d. Derive a closed form (involving d) for the generating function $E(x) ::= \sum_0^\infty e_n x^n$.

(c) Find a closed form (involving d) for e_n.

(d) Use part (c) to solve for d.

(e) Prove that $e_n = n(T - n)$.

Problems for Section 21.2

Practice Problems

Problem 21.5.
Consider the following random-walk graphs:

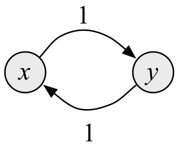


Figure 21.4

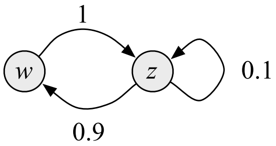


Figure 21.5

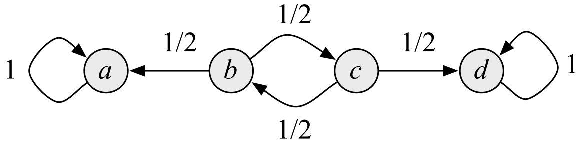


Figure 21.6

(a) Find $d(x)$ for a stationary distribution for graph 21.4.

(b) Find $d(y)$ for a stationary distribution for graph 21.4.

(c) If you start at node x in graph 21.4 and take a (long) random walk, does the distribution over nodes ever get close to the stationary distribution?

(d) Find $d(w)$ for a stationary distribution for graph 21.5.

(e) Find $d(z)$ for a stationary distribution for graph 21.5.

(f) If you start at node w in graph 21.5 and take a (long) random walk, does the distribution over nodes ever get close to the stationary distribution? (*Hint:* try a few steps and watch what is happening.)

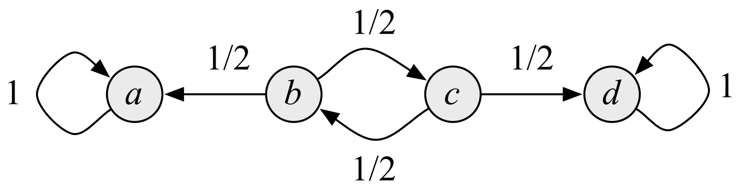
(g) How many stationary distributions are there for graph 21.6?

(h) If you start at node b in graph 21.6 and take a (long) random walk, what will be the approximate probability that you are at node d?

Problem 21.6.
A *sink* in a digraph is a vertex with no edges leaving it. Circle whichever of the following assertions are true of stable distributions on finite digraphs with exactly two sinks:

- there may not be any
- there may be a unique one
- there are exactly two
- there may be a countably infinite number
- there may be a uncountable number
- there always is an uncountable number

Problem 21.7.
Explain why there are an uncountable number of stationary distributions for the following random walk graph.



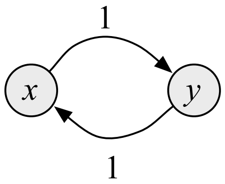


Figure 21.7

Class Problems

Problem 21.8. **(a)** Find a stationary distribution for the random walk graph in Figure 21.7.

(b) Explain why a long random walk starting at node x in Figure 21.7 will not converge to a stationary distribution. Characterize which starting distributions will converge to the stationary one.

(c) Find a stationary distribution for the random walk graph in Figure 21.8.

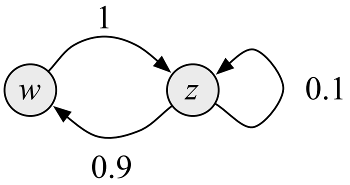


Figure 21.8

(d) If you start at node w Figure 21.8 and take a (long) random walk, does the distribution over nodes ever get close to the stationary distribution? You needn’t prove anything here, just write out a few steps and see what’s happening.

(e) Explain why the random walk graph in Figure 21.9 has an uncountable number of stationary distributions.

(f) If you start at node b in Figure 21.9 and take a long random walk, the probability you are at node d will be close to what fraction? Explain.

(g) Give an example of a random walk graph that is not strongly connected but has a unique stationary distribution. *Hint:* There is a trivial example.

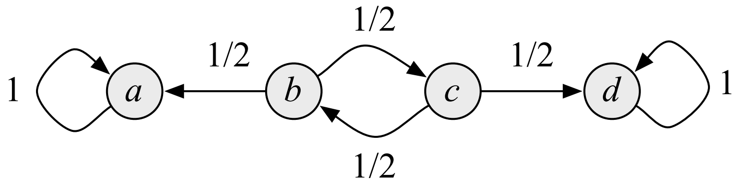


Figure 21.9

Problem 21.9.
We use random walks on a digraph G to model the typical movement pattern of a Math for CS student right after the final exam.

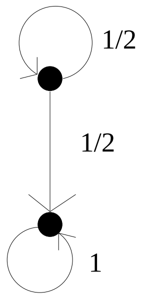
The student comes out of the final exam located on a particular node of the graph, corresponding to the exam room. What happens next is unpredictable, as the student is in a total haze. At each step of the walk, if the student is at node u at the end of the previous step, they pick one of the edges $\langle u \to v \rangle$ uniformly at random from the set of all edges directed out of u, and then walk to the node v.

Let $n ::= |V(G)|$ and define the vector $P^{(j)}$ to be

$$P^{(j)} ::= (p_1^{(j)}, \ldots, p_n^{(j)})$$

where $p_i^{(j)}$ is the probability of being at node i after j steps.

(a) We will start by looking at a simple graph. If the student starts at node 1 (the top node) in the following graph, what is $P^{(0)}$, $P^{(1)}$, $P^{(2)}$? Give a nice expression for $P^{(n)}$.



(b) Given an arbitrary graph, show how to write an expression for $p_i^{(j)}$ in terms of the $p_k^{(j-1)}$'s.

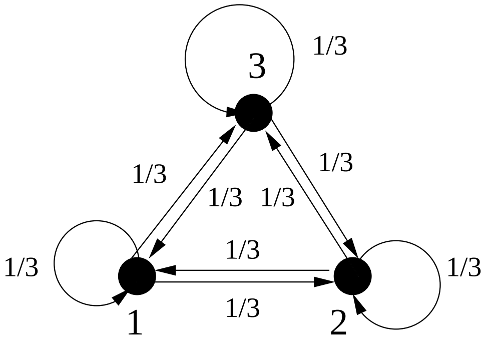
(c) Does your answer to the last part look like any other system of equations you've seen in this course?

(d) Let the *limiting distribution* vector π be

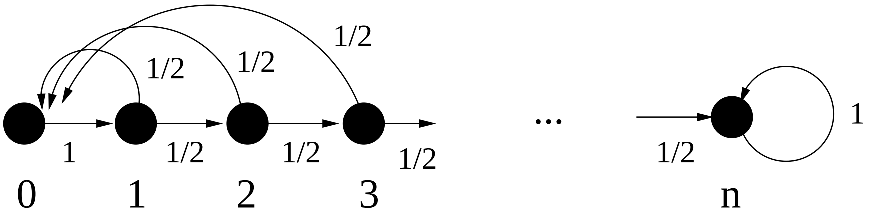
$$\lim_{k\to\infty} \frac{\sum_{i=1}^{k} P^{(i)}}{k}.$$

What is the limiting distribution of the graph from part a? Would it change if the start distribution were $P^{(0)} = (1/2, 1/2)$ or $P^{(0)} = (1/3, 2/3)$?

(e) Let’s consider another directed graph. If the student starts at node 1 with probability 1/2 and node 2 with probability 1/2, what is $P^{(0)}$, $P^{(1)}$, $P^{(2)}$ in the following graph? What is the limiting distribution?



(f) Now we are ready for the real problem. In order to make it home, the poor Math for student is faced with n doors along a long hall way. Unbeknownst to him, the door that goes outside to paradise (that is, freedom from the class and more importantly, vacation!) is at the *very end*. At each step along the way, he passes by a door which he opens up and goes through with probability 1/2. Every time he does this, he gets teleported back to the exam room. Let’s figure out how long it will take the poor guy to escape from the class. What is $P^{(0)}$, $P^{(1)}$, $P^{(2)}$? What is the limiting distribution?



(g) Show that the expected number $T(n)$ of teleportations you make back to the exam room before you escape to the outside world is $2^{n-1} - 1$.

Problem 21.10.
Prove that for finite random walk graphs, the uniform distribution is stationary iff the probabilities of the edges coming into each vertex always sum to 1, namely

$$\sum_{u \in \text{into}(v)} p(u, v) = 1, \tag{21.18}$$

where $\text{into}(w) ::= \{v \mid \langle v \to w \rangle \text{ is an edge}\}$.

Problem 21.11.
A Google-graph is a random-walk graph such that every edge leaving any given vertex has the same probability. That is, the probability of each edge $\langle v \to w \rangle$ is $1/\operatorname{outdeg}(v)$.

A digraph is *symmetric* if, whenever $\langle v \to w \rangle$ is an edge, so is $\langle w \to v \rangle$. Given any finite, symmetric Google-graph, let

$$d(v) ::= \frac{\operatorname{outdeg}(v)}{e},$$

where e is the total number of edges in the graph.

(a) If d was used for webpage ranking, how could you hack this to give your page a high rank? ...explain informally why this wouldn't work for "real" page rank with nonsymmetric digraphs.

(b) Show that d is a stationary distribution.

Homework Problems

Problem 21.12.
A digraph is *strongly connected* iff there is a directed path between every pair of distinct vertices. In this problem we consider a finite random walk graph that is strongly connected.

(a) Let d_1 and d_2 be distinct distributions for the graph, and define the *maximum dilation* γ of d_1 over d_2 to be

$$\gamma ::= \max_{x \in V} \frac{d_1(x)}{d_2(x)}.$$

Call a vertex x *dilated* if $d_1(x)/d_2(x) = \gamma$. Show that there is an edge $\langle y \to z \rangle$ from an undilated vertex y to a dilated vertex z. *Hint:* Choose any dilated vertex x and consider the set D of dilated vertices connected to x by a directed path (going to x) that only uses dilated vertices. Explain why $D \neq V$, and then use the fact that the graph is strongly connected.

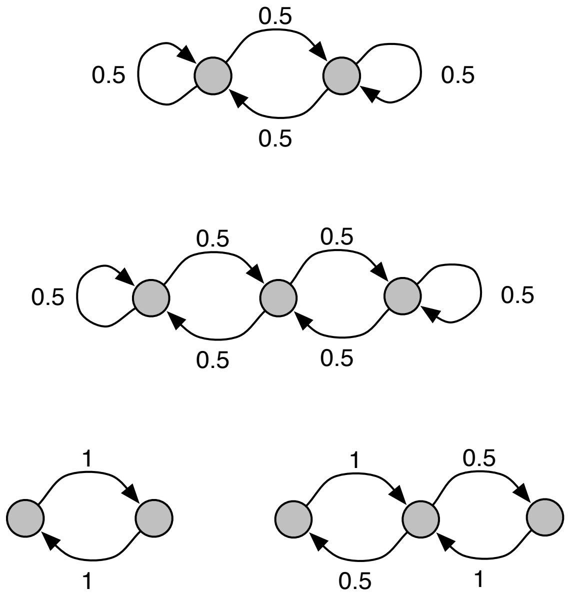


Figure 21.10 Which ones have uniform stationary distribution?

(b) Prove that the graph has *at most one* stationary distribution. (There always *is* a stationary distribution, but we're not asking you prove this.) *Hint:* Let d_1 be a stationary distribution and d_2 be a different distribution. Let z be the vertex from part (a). Show that starting from d_2, the probability of z changes at the next step. That is, $\widehat{d_2}(z) \neq d_2(z)$.

Exam Problems

Problem 21.13.
For which of the graphs in Figure 21.10 is the uniform distribution over nodes a stationary distribution? The edges are labeled with transition probabilities. Explain your reasoning.

Problem 21.14.
A *trap* in a random walk digraph is a subgraph whose edges all end within the trap. Suppose a graph has exactly two traps, and they do not share any vertices. Circle whichever of the following assertions are **True** of stable distributions of the graph.

(a) There may not be any.

(b) There is a unique one.

(c) There are exactly two.

(d) There are only a finite number.

(e) There are a countable number.

(f) In some such graphs there are a countable number, and in other such graphs there are an uncountable number.

(g) There are an uncountable number.