# 19 Random Processes

*Random Walks* are used to model situations in which an object moves in a sequence of steps in randomly chosen directions. For example in Physics, three-dimensional random walks are used 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.

## 19.1 ~~Gamblers’~~ 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, then he is called an overall *winner*, and he will have won his intended profit. If his capital reaches zero dollars before reaching his target, then we say that he is “ruined” or *goes broke*, and he will have lost $n$ dollars.~~ 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 19.1. The random walk has boundaries at 0 and $T$. If the random walk ever reaches either of these boundary values, then it terminates.

In a *fair game*, the gambler is equally likely to win or lose each bet~~—~~that is $p = 1/2$. The corresponding random walk is called *unbiased*. 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 $T$, ~~namely,~~ the probability that the gambler wins. We’ll do this in

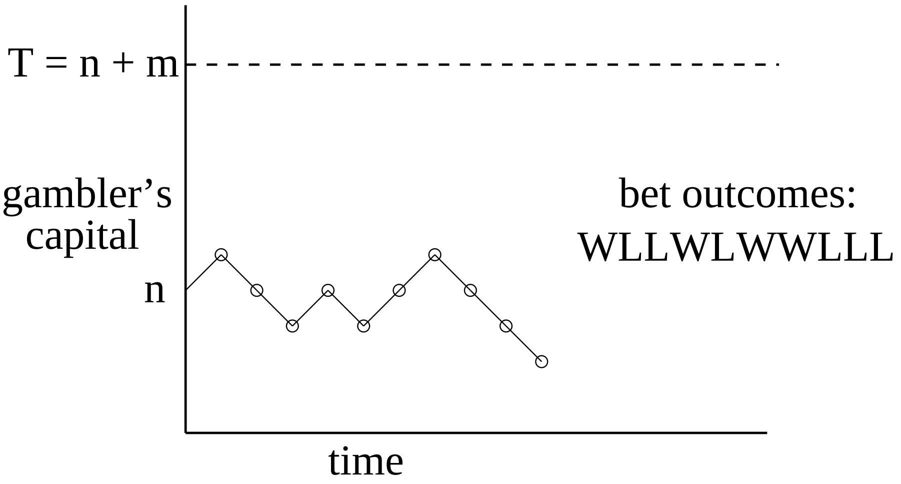


**Figure 19.1** A graph of the gambler's capital versus time for one possible sequence of bet outcomes. At each time step, the graph goes up with probability $p$ and down with probability $1 - p$. The gambler continues betting until the graph reaches either 0 or $T$. If he starts with $\$n$, his intended profit is $\$m$ where $T = n + m$.

Section 19.1.1, but before we derive the probability, let's just look at what it turns out to be.

Let's begin by supposing ~~the coin is fair, the gambler starts with 100 dollars, and he wants to double his money~~. ~~That is, he plays until he goes broke or reaches a target of 200 dollars.~~ Since he starts equidistant from his target and bankruptcy, it's clear by symmetry that his probability of winning in this case 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 fair 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, the game is still fair. When he wins, he only wins \$100~~;~~ when he loses, he loses big: \$1M. So the gambler's average win ~~is actually~~ zero dollars.

Another 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 $T = 100 + 500 = 600$. ~~So~~ 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. ~~This game is slightly biased against the gambler: the probability of winning a single bet is~~ $p = 18/38 \approx 0.47$. ~~(It’s the two green numbers that slightly bias the bets and give the casino an edge.)~~ ~~Still, the bets are almost 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.

~~Not so! The gambler’s odds of winning $100 making one dollar bets against the “slightly” unfair roulette wheel are less than 1 in 37,000. If that seems surprising, listen to this: *no matter how much money* the gambler has to start — $5000, $50,000, $5 · 10^12 — his odds are still less than 1 in 37,000 of winning a mere 100 dollars!~~

~~Moral: Don’t play!~~

~~The theory of random walks is filled with such fascinating and counter-intuitive conclusions.~~

### 19.1.1 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 of the subject of probability.

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 probabillity $p$ and loses his top chip to Eric with probabillity $q ::= 1 - p$. They play this game until one person goes bankrupt.

Pascal’s ingenious idea was to alter the value of the chips to make the game fair. ~~Namely, Albert’s~~ bottom chip will be given payoff value $r$ where $r ::= q/p$, and the successive chips *up* his stack will be worth $r^2, r^3, \ldots$ up to his top chip with payoff value $r^n$. Eric’s top chip will be worth $r^{n+1}$ and the successive chips *down* his stack will be worth $r^{n+2}, r^{n+3}, \ldots$ down to his bottom chip worth $r^{n+m}$.

~~Now the~~ expected change in Albert’s chip values on the first bet is

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

so this payoff makes the bet fair. Moreover, whether Albert wins or loses the bet, the successive chip values 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 payoff remains fair. So Albert's expected payoff at the end of the game is the sum of the expectations of his payoffs of each bet, namely 0. Here we're legitimately appealing to infinite linearity, since the payoff amounts remain bounded independent of the number of bets.

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

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

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

$$w_n = \frac{r^n - 1}{r^{n+m} - 1} = \frac{r^n - 1}{r^T - 1} \tag{19.1}$$

We have now proved

**Theorem 19.1.1.** *In the Gambler's Ruin game with initial capital, $n$, target, $T$, and probability $p$ of winning each individual bet,*

$$\Pr[\textit{the gambler wins}] = \begin{cases} \dfrac{n}{T} & \textit{for } p = \dfrac{1}{2}, \\[2ex] \dfrac{r^n - 1}{r^T - 1} & \textit{for } p \neq \dfrac{1}{2}, \end{cases} \tag{19.2}$$

*where $r ::= q/p$.*

### 19.1.2 A Recurrence for the Probability of Winning

Pascal was obviously a clever fellow, but ~~fortunately for the rest of us less ingenious folks,~~ linear recurrences offer a methodical, if less inspiring, approach to Gambler's Ruin.

The probability that the gambler wins is a function of his initial capital, $n$, his target, $T \geq n$, and the probability, $p$, that he 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

$$w_0 = 0, \tag{19.3}$$
$$w_T = 1. \tag{19.4}$$

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

$$w_{n+1} = \frac{w_n}{p} - rw_{n-1} \tag{19.5}$$

where $r$ is $q/p$ as in section 19.1.1.

This recurrence holds only for $n+1 \le T$, but there's no harm in using (19.5) to define $w_{n+1}$ for all $n+1 > 1$. Now, letting

$$W(x) ::= w_0 + w_1x + w_2x^2 + \cdots$$

be the generating function for the $w_n$, we derive from (19.5) and (19.3) using our generating function methods that

$$W(x) = \frac{w_1x}{rx^2 - x/p + 1}. \tag{19.6}$$

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

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

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

$$W(x) = \frac{A}{1-x} + \frac{B}{1-rx}, \tag{19.7}$$

for some constants $A, B$. To solve for $A, B$, note that by (19.6) and (19.7),

$$w_1x = A(1-rx) + B(1-x),$$

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

$$w_n = w_1 \frac{r^n - 1}{r - 1}. \tag{19.8}$$

Finally, we can use (19.8) 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 (19.8), we arrive at the solution:

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

matching Pascal's result (19.1).

In the unbiased case where $p = q$, we get from (19.6) that

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

and again can use partial fractions to match Pascal's result (19.2).

### A simpler expression for the biased case

The expression (19.1) 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 (19.1) 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 19.1.2.** *In the Gambler's Ruin game with initial capital, $n$, target, $T$, and probability $p < 1/2$ of winning each individual bet,*

$$\Pr[\textit{the gambler wins}] < \left(\frac{1}{r}\right)^{T-n} \tag{19.9}$$

*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 (19.9) decreases exponentially ~~with~~ the intended profit. So, for example, doubling his intended profit will square his probability of winning. In ~~particular,~~ 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.

### 19.1.3 Intuition

Why is the gambler so unlikely to make money when the game is slightly biased against him? ~~Intuitively, there are two forces at work~~. First, the gambler's capital has random upward and downward *swings* due to 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 19.2.

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. ~~The problem is that~~ 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 ~~a few hundred dollars, he needs a huge upward swing to save himself~~. ~~And such a huge swing is extremely improbable.~~ 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 we saw before that 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

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

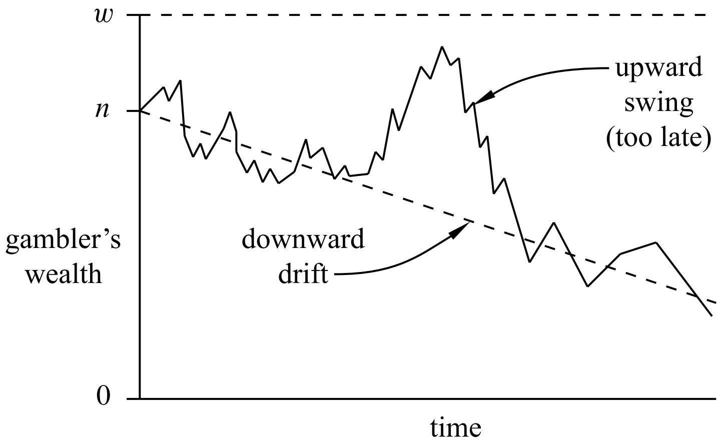


**Figure 19.2** In a biased random walk, the downward drift usually dominates swings of good luck.

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.

### 19.1.4 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,

$$\text{Ex}[e_n \mid \text{gambler wins first bet}] = 1 + e_{n+1}.$$

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

bets:

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

So we have

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

This yields the linear recurrence

$$e_{n+1} = \frac{1}{p}e_n - \frac{q}{p}e_{n-1} - \frac{1}{p}. \tag{19.10}$$

The routine solution of this linear recurrence yields:

**Theorem 19.1.3.** *In the Gambler's Ruin game with initial capital $n$, target $T$, and probability $p$ of winning each bet,*

$$\text{Ex}[\textit{number of bets}] = \begin{cases} n(T-n) & \textit{for } p = \dfrac{1}{2}, \\ \dfrac{\frac{r^n-1}{r^T-1} \cdot T - n}{p-q} & \textit{for } p \neq \dfrac{1}{2}. \end{cases} \tag{19.11}$$

In the unbiased case, (19.11) can be rephrased simply as

$$\text{Ex}[\text{number of fair bets}] = \text{initial capital} \cdot \text{intended profit}. \tag{19.12}$$

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 (19.12) cries out for an intuitive proof, but we have not found one (where are you, Pascal?).

### 19.1.5 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, $p \leq 1/2$, the gambler is sure to go broke. ~~In particular, even in a "fair" game with $p = 1/2$, he is sure to go broke.~~

**Lemma 19.1.4.** *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 19.1.5.** *If the gambler starts with one or more dollars and plays a fair unbounded game, then his expected number of plays is infinite.*

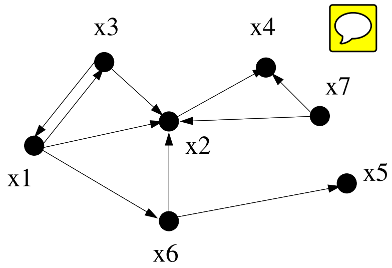
A proof appears in Problem 19.2.

So even starting with just one dollar, the expected number of plays before going broke is infinite! ~~Of course, this does not mean~~ that the gambler is *likely* to play for long —there is even a 50% chance he will lose the very first bet and go broke right away.

Lemma 19.1.5 says that the gambler can "expect" to play forever, while Lemma 19.1.4 says that he is certain to go broke. These facts sound contradictory, but they are sound consequences of the technical mathematical definition of expectation. The moral here, as in section 18.8, is that naive intuition is unreliable when it comes to infinite expectation.

## 19.2 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 $y$. For example, in the following graph the vertices $x_1, \ldots, x_n$ correspond to web pages and $\langle x_i \to x_j \rangle$ is a directed edge when page $x_i$ contains a hyperlink to page $x_j$.



The web graph is an enormous graph with ~~many billions and probably even~~ trillions of vertices. At first glance, this graph wouldn’t seem to be very interesting. But 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. ~~So if an author wanted a document to get a higher score for certain keywords, he would put the keywords in the title and make it appear in lots of places. You can even see this today with some bogus web sites.~~

This approach works fine if you only have a few documents that match a search term. But on the web, there are 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! How does Google decide which 10 or 20 to show first? It wouldn’t be smart to pick a page that gets a high keyword score because it has “Mathematics Mathematics . . . Mathematics” across the front of the document.

~~One way to get placed high on the list is to pay Google an advertising fee —and Google gets an enormous revenue stream from these fees. Of course an early listing is worth a fee only if an advertiser’s target audience is attracted to the listing. But an audience does get attracted to Google listings because its ranking method is really good at determining the most relevant web pages,~~ For example, Google demonstrated its accuracy in our case by giving first rank to our 6.042 text[1] ~~: -)~~ . So how did Google know to pick 6.042 to be first out of 482,000?

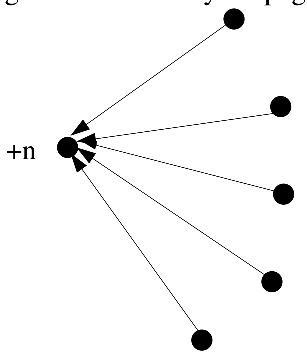
Well back in 1995, Larry and Sergey got the idea to allow the digraph structure of the web to determine which pages are likely to be the most important.

[1] 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.

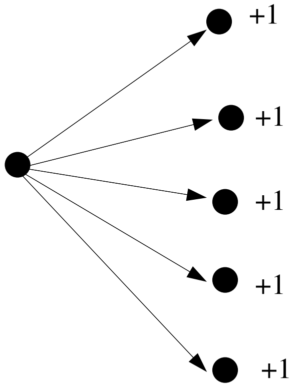
### 19.2.1 A First Crack at Page Rank

Looking at the web graph, any idea which vertex/page might be the best to rank 1st? 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 the number of links pointing to $x$, ~~that is,~~ indegree$(x)$. The idea is to think of web pages as voting for the most important page —the more votes, the better rank.

~~Of course,~~ 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.



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.

### 19.2.2 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. That is, they assigned each edge $x \to y$ of the web graph with a probability conditioned on being on page $x$:

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

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

~~For example, if the user is at page $x$, and there are three links from page $x$, then each link is followed with probability $1/3$.~~

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

$$\begin{aligned}\Pr[\text{go to } y] &= \sum_{\text{edges } \langle x \to y \rangle} \Pr\left[\text{follow link } \langle x \to y \rangle \mid \text{at page } x\right] \cdot \Pr[\text{at page } x] \\ &= \sum_{\text{edges } \langle x \to y \rangle} \frac{\Pr[\text{at } x]}{\text{outdegree}(x)} \qquad (19.13)\end{aligned}$$

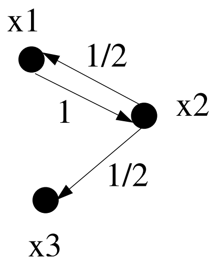
For example, in our web graph, we have

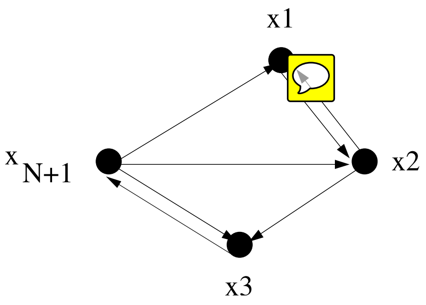
$$\Pr[\text{go to } x_4] = \frac{\Pr[\text{at } x_7]}{2} + \frac{\Pr[\text{at } x_2]}{1}.$$

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.

To model this aspect of the web, Sergey and Larry added a supervertex to the web graph and had every page with no hyperlinks point to it. Moreover, the supervertex points to every other vertex in the graph, allowing you to restart the walk from a random place. For example, below left is a graph and below right is the same graph after adding the supervertex $x_{N+1}$.





The addition of the supervertex also removes the possibility that the value $1/\text{outdegree}(x)$ might involve a division by zero.

### 19.2.3 Stationary Distribution & Page Rank

The basic idea of page rank is just 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

$$\sum_{\text{vertices } x} \Pr[\text{at } x] = 1. \tag{19.14}$$

**Definition 19.2.1.** 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, $\text{PR}(x)$, for each vertex, $x$, such that

$$\text{PR}(x) = \sum_{\text{edges } \langle y \rightarrow x \rangle} \frac{\text{PR}(y)}{\text{outdegree}(y)}, \tag{19.15}$$

corresponding to the intuitive equations given in (19.13). These numbers must also satisfy the additional constraint corresponding to (19.14):

$$\sum_{\text{vertices } x} \text{PR}(x) = 1. \tag{19.16}$$

So if there are $n$ vertices, then equations (19.15) and (19.16) provide a system of $n + 1$ linear equations in the $n$ variables, $\text{PR}(x)$. Note that constraint (19.16) is needed because the remaining constraints (19.15) could be satisfied by letting $\text{PR}(x) ::= 0$ 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. Their addition of a supervertex ensures there is always a *unique* stationary distribution. 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. Examples of this appear in some of the problems below.

Now just keeping track of the digraph whose vertices are billions of web pages is a daunting task. That's why Google is building power plants. 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 6.042 ranked first out of 378,000 matches. Lots of other universities used our notes and presumably have links to the 6.042 open courseware site, and the university sites themselves are legitimate, which ultimately leads to 6.042 getting a high page rank in the web graph.

## Problems for Section 19.1

### Practice Problems

**Problem 19.1.**
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 19.2.**
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 19.3.**
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 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 for $e_n$; you need *not* solve the recurrence.

**Problem 19.4.**
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

$$e_n = ae_{n-1} + be_{n-2} + c. \tag{19.17}$$

for $1 < n < T$.

**(b)** Let $e_n$ be defined by (19.17) 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 $e_n = n(T - n)$.

a

## Problems for Section 19.2

### Practice Problems

**Problem 19.5.**
Consider the following random-walk graphs:

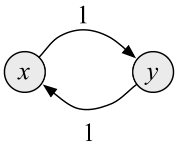


**Figure 19.3**

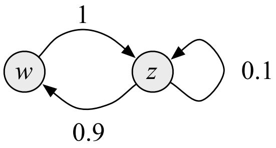


**Figure 19.4**

**(a)** Find $d(x)$ for a stationary distribution for graph 19.3.

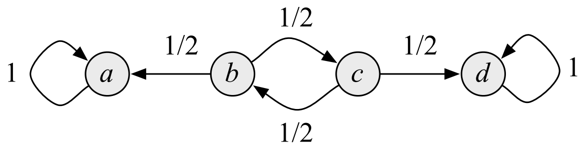


**Figure 19.5**

**(b)** Find $d(y)$ for a stationary distribution for graph 19.3.

**(c)** If you start at node $x$ in graph 19.3 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 19.4.

**(e)** Find $d(z)$ for a stationary distribution for graph 19.4.

**(f)** If you start at node $w$ in graph 19.4 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.)

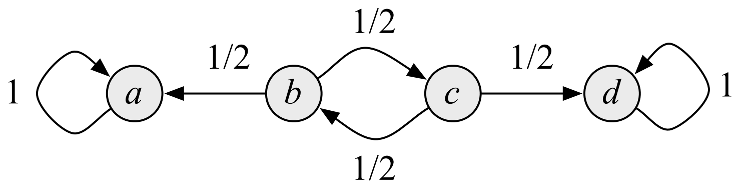
**(g)** How many stationary distributions are there for graph 19.5?

**(h)** If you start at node $b$ in graph 19.5 and take a (long) random walk, what will be the approximate probability that you are at node $d$?

**Problem 19.6.**
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 19.7.**
Explain why there are an uncountable number of stationary distributions for the following random walk graph.



### Class Problems

**Problem 19.8. (a)** Find a stationary distribution for the random walk graph in Figure 19.6.

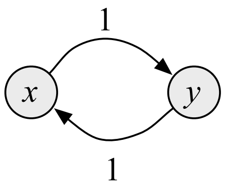


**Figure 19.6**

**(b)** If you start at node $x$ in Figure 19.6 and take a (long) random walk, does the distribution over nodes ever get close to the stationary distribution? Explain.

**(c)** Find a stationary distribution for the random walk graph in Figure 19.7.

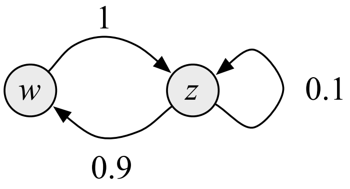


**Figure 19.7**

**(d)** If you start at node $w$ Figure 19.7 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.

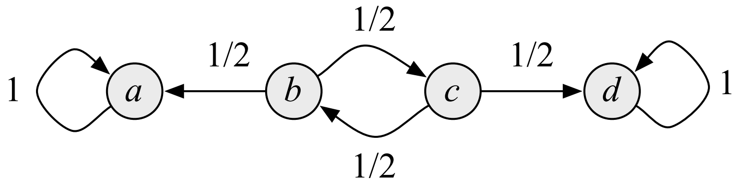


**Figure 19.8**

**(e)** Find a stationary distribution for the random walk graph in Figure 19.8.

**(f)** If you start at node $b$ in Figure 19.8 and take a long random walk, the probability you are at node $d$ will be close to what fraction? Explain.

**Problem 19.9.**
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.

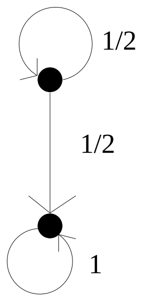
The student comes out of the final exam located on a particular node of the graph, corresponding to the exam room. What happens next is unpredictable, as the student is in a total haze. At each step of the walk, if the student is at node $u$ at the end of the previous step, they pick one of the edges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

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

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.

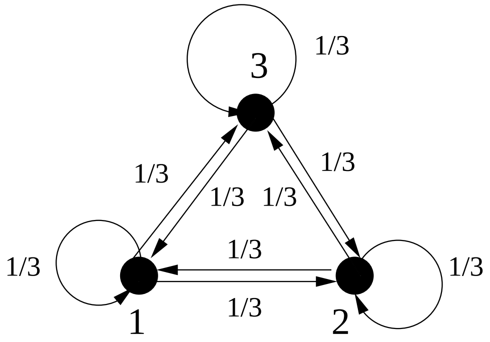
**(c)** Does your answer to the last part look like any other system of equations you've seen in this course?

**(d)** Let the *limiting distribution* vector, $\pi$, be

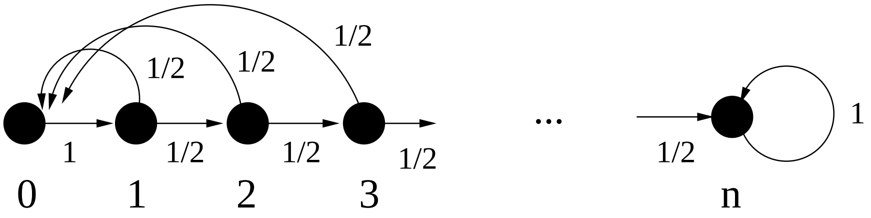
$$\lim_{k\to\infty} \frac{\sum_{i=1}^{k} P^{(i)}}{k}.$$

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 19.10.**
Prove that for finite random walk graphs, the uniform distribution is stationary if and only the probabilities of the edges coming into each vertex always sum to 1, namely

$$\sum_{u \in \text{into}(v)} p(u, v) = 1, \tag{19.18}$$

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

**Problem 19.11.**
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? ...and explain informally why this wouldn't work for "real" page rank using digraphs?

**(b)** Show that $d$ is a stationary distribution.

### Homework Problems

**Problem 19.12.**
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*, $\gamma$, of $d_1$ over $d_2$ to be

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

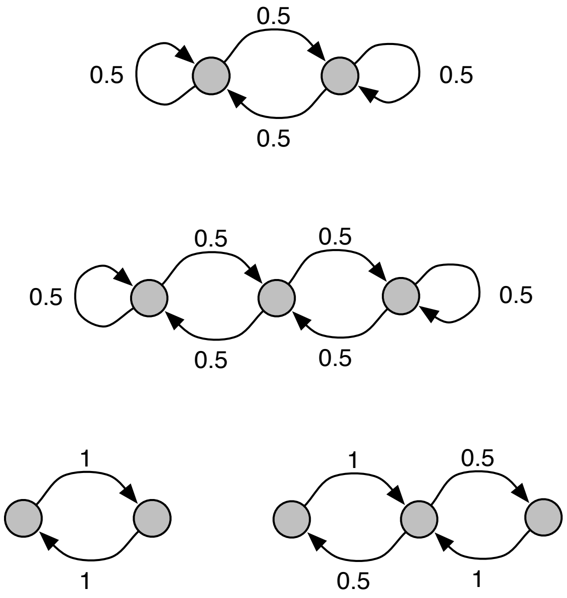


**Figure 19.9** Which ones have uniform stationary distribution?

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.

**(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 19.13.**
For which of the graphs in Figure 19.9 is the uniform distribution over nodes a stationary distribution? The edges are labeled with transition probabilities. Explain your reasoning.

# Index