Quiz

- For an $n \times n$ matrix A, define what it means for something to be an eigenvector and and what it means for something to be an eigenvalue of A. (Is the zero vector an eigenvector?)
- What does it mean if 0 is an eigenvalue of a matrix A?

Not for credit

Suppose A is an $n \times n$ matrix, and suppose $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are eigenvectors that form a basis for $\mathbb{R}^n$. Let $\mathbf{v}$ be a vector in $\mathbb{R}^n$. Suppose that the coordinate representation of $\mathbf{v}$ in terms of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is $[\alpha_1, \ldots, \alpha_n]$. What is the coordinate representation of $A\mathbf{v}$?

Interpretation using change of basis, revisited

A diagonalizable $\Rightarrow A = S\Lambda S^{-1}$ for a diag. Λ and invertible S.
Suppose $\mathbf{x}^{(0)}$ is a vector. The equation $\mathbf{x}^{(t+1)} = A\,\mathbf{x}^{(t)}$ then defines $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \ldots$. Then

$$\begin{aligned}\mathbf{x}^{(t)} &= \underbrace{A\,A\,\cdots\,A}_{t \text{ times}}\,\mathbf{x}^{(0)} \\ &= (S\Lambda S^{-1})(S\Lambda S^{-1})\cdots(S\Lambda S^{-1})\,\mathbf{x}^{(0)} \\ &= S\Lambda^t S^{-1}\,\mathbf{x}^{(0)}\end{aligned}$$

Interpretation: Let $\mathbf{u}^{(t)}$ = coordinate representation of $\mathbf{x}^{(t)}$ in terms of the columns of S. Then we have the equation $\mathbf{u}^{(t+1)} = \Lambda\,\mathbf{u}^{(t)}$. Therefore

$$\begin{aligned}\mathbf{u}^{(t)} &= \underbrace{\Lambda\,\Lambda\,\cdots\,\Lambda}_{t \text{ times}}\,\mathbf{u}^{(0)} \\ &= \Lambda^t\,\mathbf{u}^{(0)}\end{aligned}$$

If $\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$ then

$$\Lambda^t = \begin{bmatrix} \lambda_1^t & & \\ & \ddots & \\ & & \lambda_n^t \end{bmatrix}$$

Interpretation using change of basis, re-revisited

Suppose $n \times n$ matrix A is diagonalizable, so it has linearly independent e-vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ with e-values are $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Any vector $\mathbf{x}$ can be written as a linear combination:

$$\mathbf{x} = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n$$

Left-multiply by A on both sides of the equation:

$$\begin{aligned} A\mathbf{x} &= A(\alpha_1 \mathbf{v}_1) + A(\alpha_2 \mathbf{v}_2) + \cdots + A(\alpha_n \mathbf{v}_n) \\ &= \alpha_1 A\mathbf{v}_1 + \alpha_2 A\mathbf{v}_2 + \cdots + \alpha_n A\mathbf{v}_n \\ &= \alpha_1 \lambda_1 \mathbf{v}_1 + \alpha_2 \lambda_2 \mathbf{v}_2 + \cdots + \alpha_n \lambda_n \mathbf{v}_n \end{aligned}$$

Applying the same reasoning to $A(A\mathbf{x})$, we get

$$A^2\mathbf{x} = \alpha_1 \lambda_1^2 \mathbf{v}_1 + \alpha_2 \lambda_2^2 \mathbf{v}_2 + \cdots + \alpha_n \lambda_n^2 \mathbf{v}_n$$

More generally, for any nonnegative integer t,

$$A^t\mathbf{x} = \alpha_1 \lambda_1^t \mathbf{v}_1 + \alpha_2 \lambda_2^t \mathbf{v}_2 + \cdots + \alpha_n \lambda_n^t \mathbf{v}_n$$

If $|\lambda_1| > |\lambda_2|$ then eventually λ_1^t will be *much* bigger than $\lambda_2^t, \ldots, \lambda_n^t$, so first term will dominate. For a large enough value of t, $A^t\mathbf{x}$ will be approximately $\alpha_1 \lambda_1^t \mathbf{v}_1$.

Rabbit reproduction and death

A disease enters the rabbit population. In each month,

- A δ fraction of the adult population catches it,
- an ϵ fraction of the juvenile population catches it, and
- an η fraction of the sick population die, and the rest recover.

Sick rabbits don't produce babies.

Equations

$$\begin{array}{rcccccc} \text{well adults}' & = & (1-\delta)\text{ well adults} & + & (1-\epsilon)\text{juveniles} & + & (1-\eta)\text{ sick} \\ \text{juveniles}' & = & \text{well adults} & & & & \\ \text{sick}' & = & \delta\text{ well adults} & + & \epsilon\text{ juveniles} & + & 0 \end{array}$$

Represent change in populations by matrix-vector equation

$$\begin{bmatrix} \text{well adults at time } t+1 \\ \text{juveniles at time } t+1 \\ \text{sick at time } t+1 \end{bmatrix} = \begin{bmatrix} 1-\delta & 1-\epsilon & (1-\eta) \\ 1 & 0 & 0 \\ \delta & \epsilon & 0 \end{bmatrix} \begin{bmatrix} \text{well adults at time } t \\ \text{juveniles at time } t \\ \text{sick at time } t \end{bmatrix}$$

(You might question fractional rabbits, deterministic infection.)

Question: Does the rabbit population still grow?

Analyzing the rabbit population in the presence of disease

$$\begin{bmatrix} \text{well adults at time } t+1 \\ \text{juveniles at time } t+1 \\ \text{sick at time } t+1 \end{bmatrix} = \begin{bmatrix} 1-\delta & 1-\epsilon & (1-\eta) \\ 1 & 0 & 0 \\ \delta & \epsilon & 0 \end{bmatrix} \begin{bmatrix} \text{well adults at time } t \\ \text{juveniles at time } t \\ \text{sick at time } t \end{bmatrix}$$

Question: Does the rabbit population still grow?
Depends on the values of the parameters ($\delta =$ infection rate among adults, $\epsilon =$ infection rate among juveniles, $\eta =$ death rate among sick).

Plug in different values for parameters and then compute eigenvalues

- $\delta = 0.5, \epsilon = 0.5, \eta = 0.8$. The largest eigenvalue is 1.1172 (with eigenvector $[0.6299, 0.5638, 0.5342]$). This means that the population grows exponentially in time (roughly proportional to 1.117^t).
- $\delta = 0.7, \epsilon = 0.7, \eta = 0.8$. The largest eigenvalue is 0.9327. This means the population shrinks exponentially.
- $\delta = 0.6, \epsilon = 0.6, \eta = 0.8$. The largest eigenvalue is 1.02. Population grows exponentially.

Expected number of rabbits

There's these issues of fractional rabbits and deterministic disease.
The matrix-vector equation really describes the *expected values* of the various populations.

Modeling population movement

Dance-club dynamics: At the beginning of each song,

- 56% of the people standing on the side go onto the dance floor, and
- 12% of the people on the dance floor leave it.

Suppose that there are a hundred people in the club. Assume nobody enters the club and nobody leaves. What happens to the number of people in each of the two locations?

Represent state of system by

$$\mathbf{x}^{(t)} = \begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \end{bmatrix} = \begin{bmatrix} \text{number of people standing on side after } t \text{ songs} \\ \text{number of people on dance floor after } t \text{ songs} \end{bmatrix}$$

$$\begin{bmatrix} x_1^{(t+1)} \\ x_2^{(t+1)} \end{bmatrix} = \begin{bmatrix} .44 & .12 \\ .56 & .88 \end{bmatrix} \begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \end{bmatrix}$$

Diagonalize: $S^{-1}AS = \Lambda$ where

$$A = \begin{bmatrix} .44 & .12 \\ .56 & .88 \end{bmatrix}, S = \begin{bmatrix} 0.209529 & -1 \\ 0.977802 & 1 \end{bmatrix}, \Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 0.32 \end{bmatrix}$$

Analyzing dance-floor dynamics

$$
\begin{aligned}
\begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \end{bmatrix} &= \left(S\Lambda S^{-1}\right)^t \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} \\
&= S\Lambda^t S^{-1} \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} \\
&= \begin{bmatrix} .21 & -1 \\ .98 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & .32 \end{bmatrix}^t \begin{bmatrix} .84 & .84 \\ -.82 & .18 \end{bmatrix} \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} \\
&= \begin{bmatrix} .21 & -1 \\ .98 & 1 \end{bmatrix} \begin{bmatrix} 1^t & 0 \\ 0 & .32^t \end{bmatrix} \begin{bmatrix} .84 & .84 \\ -.82 & .18 \end{bmatrix} \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} \\
&= 1^t(.84x_1^{(0)} + .84x_2^{(0)}) \begin{bmatrix} .21 \\ .98 \end{bmatrix} + (0.32)^t(-.82x_1^{(0)} + .18x_2^{(0)}) \begin{bmatrix} -1 \\ 1 \end{bmatrix} \\
&= 1^t \underbrace{\left(x_1^{(0)} + x_2^{(0)}\right)}_{\text{total population}} \begin{bmatrix} .18 \\ .82 \end{bmatrix} + (0.32)^t \left(-.82x_1^{(0)} + .18x_2^{(0)}\right) \begin{bmatrix} -1 \\ 1 \end{bmatrix}
\end{aligned}
$$

Analyzing dance-floor dynamics, continued

$$\begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \end{bmatrix} = \underbrace{\left(x_1^{(0)} + x_2^{(0)}\right)}_{\text{total population}} \begin{bmatrix} .18 \\ .82 \end{bmatrix} + (0.32)^t \left(-.82x_1^{(0)} + .18x_2^{(0)}\right) \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

The numbers of people in the two locations after t songs depend on the *initial* numbers of people in the two locations.

However, the dependency grows weaker as the number of songs increases: $(0.32)^t$ gets smaller and smaller, so the second term in the sum matters less and less.

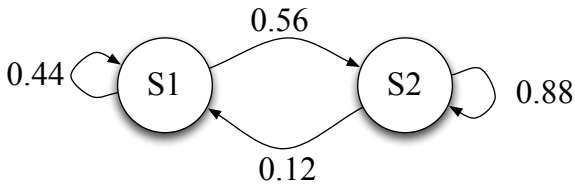
After ten songs, $(0.32)^t$ is about 0.00001.

The first term in the sum is $\begin{bmatrix} .18 \\ .82 \end{bmatrix}$ times the total number of people. This shows that, as the number of songs increases, the proportion of people on the dance floor gets closer and closer to 82%.

Modeling Randy

Without changing math, we switch interpretations. Instead of modeling whole dance-club population, we model one person, Randy.

Randy's behavior captured in transition diagram:



State S1 represents Randy being on the side. State S2 represents Randy being on the dance floor.

After each song, Randy follows one of the arrows from current state. Which arrow? Chosen randomly according to probabilities on the arrows (*transition probabilities*).

For each state, labels on arrows from that state must sum to 1.

Where is Randy?

Knowing where Randy starts at time 0 doesn't let us predict with certainty where he will be at time t. However, for each time t, we can calculate the *probability distribution* for his location.

Since there are two possible locations (off floor, on floor), the probability distribution is given by a 2-vector $\mathbf{x}^{(t)} = \begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \end{bmatrix}$ where $x_1^{(t)} + x_2^{(t)} = 1$.

Probability distribution for Randy's location at time $t+1$ is related to probability distribution for Randy's location at time t:

$$\begin{bmatrix} x_1^{(t+1)} \\ x_2^{(t+1)} \end{bmatrix} = \begin{bmatrix} .44 & .12 \\ .56 & .88 \end{bmatrix} \begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \end{bmatrix}$$

Using earlier analysis,

$$\begin{aligned} \begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \end{bmatrix} &= \left(x_1^{(0)} + x_2^{(0)}\right) \begin{bmatrix} .18 \\ .82 \end{bmatrix} + (0.32)^t \left(-.82x_1^{(0)} + .18x_2^{(0)}\right) \begin{bmatrix} -1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} .18 \\ .82 \end{bmatrix} + (0.32)^t \left(-.82x_1^{(0)} + .18x_2^{(0)}\right) \begin{bmatrix} -1 \\ 1 \end{bmatrix} \end{aligned}$$

Where is Randy?

$$\begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \end{bmatrix} = \begin{bmatrix} .18 \\ .82 \end{bmatrix} + (0.32)^t \left(-.82x_1^{(0)} + .18x_2^{(0)} \right) \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

If we know Randy starts off the dance floor at time 0 then $x_1^{(0)} = 1$ and $x_2^{(0)} = 0$.
If we know Randy starts on the dance floor at time 0 then $x_1^{(0)} = 0$ and $x_2^{(0)} = 1$.

In either case, we can plug in to equation to get exact probability distribution for time t.

But after a few songs, the starting location doesn't matter much—the probability distribution gets very close to $\begin{bmatrix} .18 \\ .82 \end{bmatrix}$ in either case.

This is called Randy's *stationary distribution*.

It doesn't mean Randy stays in one place—we expect him to move back and forth all the time.
It means that the probability distribution for his location after t steps depends less and less on t.

From Randy to spatial locality in CPU memory fetches

We again switch interpretations without changing the math.

CPU uses caches and prefetching to improve performance.
To help computer architects, it is useful to model CPU access patterns.

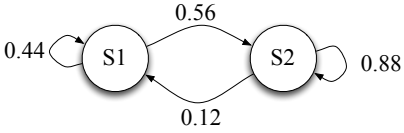
After accessing location x, CPU usually accesses location $x + 1$.
Therefore simple model is:

$$\text{Probability[address requested at time } t+1 \text{ is } 1 + \text{address requested at time } t] = .6$$

However, a slightly more sophisticated model predicts much more accurately.

Observation: Once consecutive addresses have been requested in timesteps t and $t + 1$, it is very likely that the address requested in timestep $t + 2$ is also consecutive.

Use same model as used for Randy.

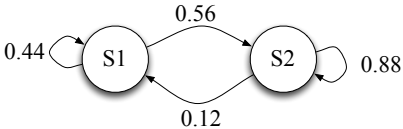


State S1 = CPU is requesting nonconsecutive addresses.
State S2 = CPU is requesting consecutive addresses.

From Randy to spatial locality in CPU memory fetches

Observation: Once consecutive addresses have been requested in timesteps t and $t+1$, it is very likely that the address requested in timestep $t+2$ is also consecutive.

Use same model as used for Randy.



State S1 = CPU is requesting nonconsecutive addresses.
State S2 = CPU is requesting consecutive addresses.

Once CPU starts requesting consecutive addresses, it tends to stay in that mode for a while. This tendency is captured by the model.

As with Randy, after a while the probability distribution is $[0.18, 0.82]$. Being in the first state means the CPU is issuing the first of a run of consecutive addresses (possibly of length 1) Since the system is in the first state roughly 18% of the time, the average length of such a run is $1/0.18$.

Various such calculations can be useful in designing architectures and improving performance.

Markov chains

An *n-state Markov chain* is a system such that

- At each time, the system is in one of n states, say $1, \ldots, n$, and
- there is a matrix A such that, if at some time t the system is in state j then for $i = 1, \ldots, n$, the probability that the system is in state i at time $t + 1$ is $A[i, j]$.

That is, $A[i, j]$ is the probability of transitioning from j to i, the $j \to i$ *transition probability*.

A is called the *transition matrix* of the Markov chain.

$A[1, 1] + A[2, 1] + \cdots + A[n, 1] = \text{Probability}(1 \to 1) + \text{Probability}(1 \to 2) + \cdots + \text{Probability}(1 \to n) = 1$

Similarly, every column's elements must sum to 1.

Called a *left stochastic matrix* (common convention is to use right stochastic matrices, where every row's elements sum to 1).

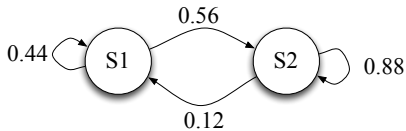
Example: $\begin{bmatrix} .44 & .12 \\ .56 & .88 \end{bmatrix}$ is the transition matrix for a two-state Markov chain.

A stationary distribution in a Markov chain

Let A = left stochastic matrix of a Markov chain: $A[i, j]$ = probability of transitioning $j \longrightarrow i$

Definition: A probability vector $\mathbf{p}$ is a *stationary distribution* for this Markov chain if $A\mathbf{p} = \mathbf{p}$. That is, $\mathbf{p}$ is an eigenvector of A corresponding to the eigenvalue 1.

Example:



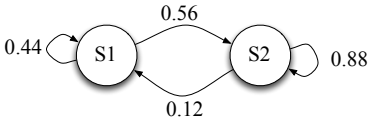
$$A = \begin{bmatrix} .44 & .12 \\ .56 & .88 \end{bmatrix}$$

$\mathbf{p} = [.18, .82] \Rightarrow A\mathbf{p} = \mathbf{p}$, so $\mathbf{p}$ is a stationary distribution for this Markov chain.

Questions:

- Are there any other stationary distributions?
- We saw that in this case the distribution gets closer and closer to this distribution. Does this happen for every Markov chain?
- How can we compute a stationary distribution?

Multiple stationary distributions

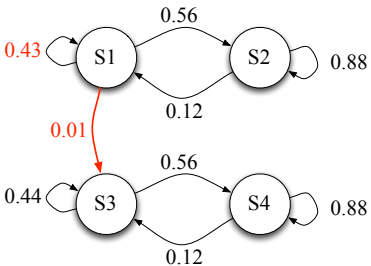


Two stationary distributions:

S1	S2	S3	S4
.18	.82	0	0

and

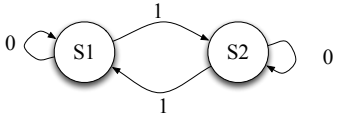
S1	S2	S3	S4
0	0	.18	.82



Back to only one stationary distribution:

S1	S2	S3	S4
0	0	.18	.82

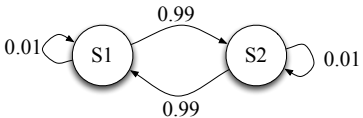
Converge to stationary distribution?



$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Does not converge to stationary distribution:

$$\mathbf{p} = \begin{array}{cc} \text{S1} & \text{S2} \\ \hline 0 & 1 \end{array} \Rightarrow A\mathbf{p} = \begin{array}{cc} \text{S1} & \text{S2} \\ \hline 1 & 0 \end{array} \text{ and vice versa}$$



$$A = \begin{bmatrix} 0.01 & 0.99 \\ 0.99 & 0.01 \end{bmatrix}$$

This one converges to stationary distribution:

S1	S2
0.5	0.5

Big Markov chains

Of course, bigger Markov chains can be useful... or fun.
A text such as a Shakespeare play can give rise to a Markov chain.
The Markov chain has one state for each word in the text.
To compute the transition probability from *word1* to *word2*, see how often an occurence of *word1* is followed immediately by *word2* (versus being followed by some other word).

Once you have constructed the transition matrix from a text, you can use it to generate random texts that resemble the original.

Or, as Zarf did, you can combine two texts to form a single text, and then generate a random text from this chimera.

Example from *Hamlet/Alice in Wonderland*:
"Oh, you foolish Alice!" she answered herself.
"How can you learn lessons in the world were now but to follow him thither with modesty enough, and likelihood to lead it, as our statists do, A baseness to write this down on the trumpet, and called out "First witness!" ... HORATIO: Most like. It harrows me with leaping in her hand, watching the setting sun, and thinking of little pebbles came rattling in at the door that led into a small passage, not much larger than a pig, my dear," said Alice (she was so

Power method

The most efficient methods for computing eigenvalues and eigenvectors are beyond scope of class.

Simple method to *sometimes* get a rough estimate of the eigenvalue of largest absolute value (and corresponding eigenvector):

Assume A is diagonalizable, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. Recall

$$A^t\mathbf{x} = \alpha_1\lambda_1^t\mathbf{v}_1 + \alpha_2\lambda_2^t\mathbf{v}_2 + \cdots + \alpha_n\lambda_n^t\mathbf{v}_n$$

If $|\lambda_1| > |\lambda_2|, \ldots, |\lambda_n|$ then first term will dominate others.

- Start with a vector $\mathbf{x}_0$.
- Find $\mathbf{x}_t = A^t\mathbf{x}_0$ by repeated matrix-vector multiplication.
- Maybe $\mathbf{x}_t$ is an approximate eigenvector corresponding to eigenvalue of largest absolute value.

Which vector $\mathbf{x}_0$ to start with? Algorithm depends on projection onto $\mathbf{v}_1$ being not too small. Random start vector should work okay. Probably all-ones vector will work too.

Power method

Failure modes of the power method:

- Initial vector might have tiny projection onto $\mathbf{v}_1$. Not likely.
- First few eigenvalues might be the same. Algorithm will "work" anyway.
- First eigenvalue might not be much bigger than next. Can still get a good estimate.
- First few eigenvalues might be different but have same absolute value. This is a real problem!

Matrix $\begin{bmatrix} 2 & & \\ & -2 & \\ & & 1 \end{bmatrix}$ has two eigenvalues with absolute value 2.

Matrix $\begin{bmatrix} \frac{1}{2} & \frac{1}{4} \\ -3 & \frac{1}{2} \end{bmatrix}$ has two complex eigenvalues, $\frac{1}{2} - \frac{\sqrt{3}}{2}\mathbf{i}$ and $\frac{1}{2} + \frac{\sqrt{3}}{2}\mathbf{i}$.

Power method applied to Markov chain

Suppose M is a Markov chain that converges to a stationary distribution from any initial distribution.
Let A be the left stochastic matrix of M.
Use power method to estimate eigenvector corresponding to eigenvalue 1

Perron-Frobenius Theorem

Theorem: Let A be an endomorphic matrix whose entries are all positive real numbers. Then

- there is only one eigenvalue of largest absolute value, and it is real; and
- it corresponds to an eigenvector whose entries are positive real numbers.

Implication...

Theorem: Let M be a Markov chain whose left stochastic matrix has no zeroes. Then M has a single stationary distribution, and it converges to this when started at any distribution.

Power method will work.

The biggest Markov chain in the world

Randy's web-surfing behavior: From whatever page he's viewing, he selects one of the links uniformly at random and follows it.

Defines a Markov chain in which the states are web pages.

Idea: Suppose this Markov chain has a stationary distribution.

- Find the stationary distribution $\Rightarrow$ probabilities for all web pages.
- Use each web page's probability as a measure of the page's importance.
- When someone searches for "matrix book", which page to return? Among all pages with those terms, return the one with highest probability.

Advantages:

- Computation of stationary distribution is independent of search terms: can be done once and subsequently used for all searches.
- Potentially could use power method to compute stationary distribution.

Pitfalls: Maybe there are several, and how would you compute one?

Using Perron-Frobenius Theorem

If can get from every state to every other state in one step, Perron-Frobenius Theorem ensures that there is only one stationary distribution....

and that the Markov chain converges to it ....

so can use power method to estimate it.
Pitfall: This isn't true for the web!

Workaround: Solve the problem with a hack: In each step, with probability 0.15, Randy just teleports to a web page chosen uniformly at random.