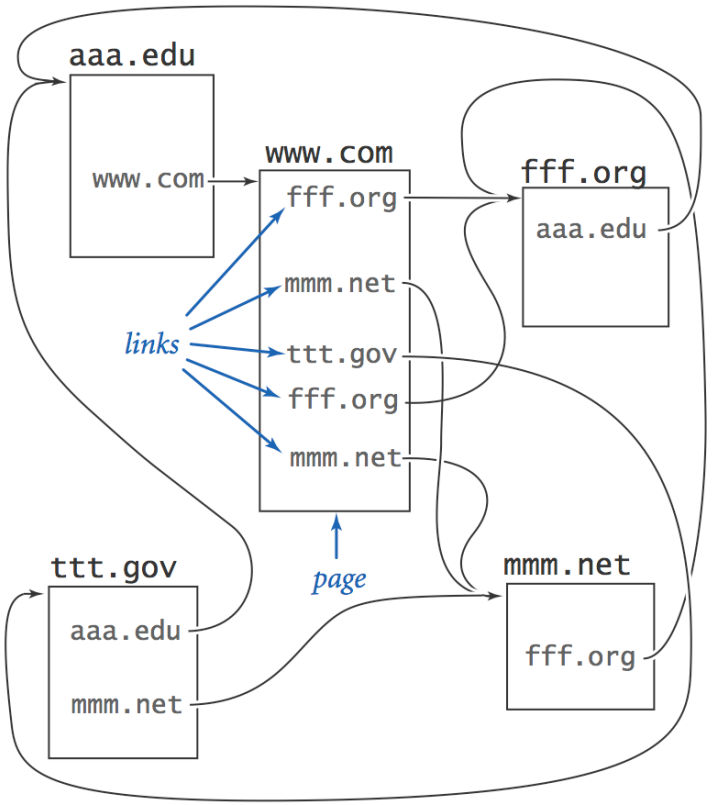
PageRank

Google's PageRank™ algorithm. [Sergey Brin and Larry Page, 1998]

- Measure popularity of pages based on hyperlink structure of Web. Revolutionized access to world's information.



90-10 Rule

Model. Web surfer chooses next page:

- 90% of the time surfer clicks random hyperlink.
- 10% of the time surfer types a random page.

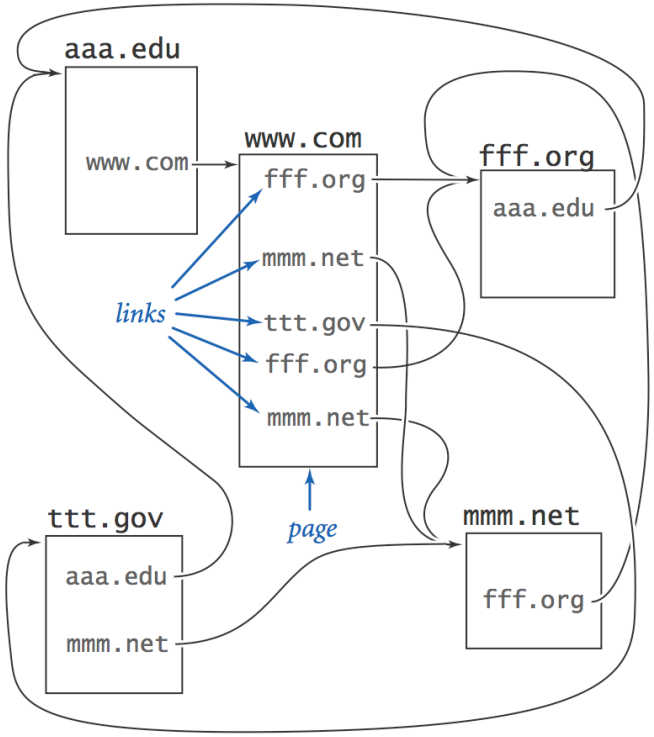
Caveat. Crude, but useful, web surfing model.

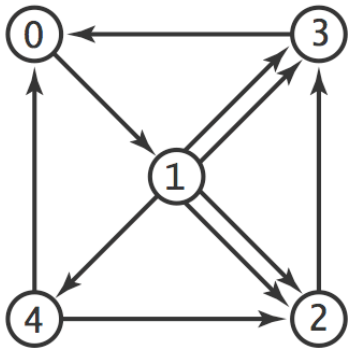
- No one chooses links with equal probability.
- No real potential to surf directly to each page on the web.
- The 90-10 breakdown is just a guess.
- It does not take the back button or bookmarks into account.
- We can only afford to work with a small sample of the web.
- …

Web Graph Input Format

Input format.

- N pages numbered 0 through N-1.
- Represent each hyperlink with a pair of integers.



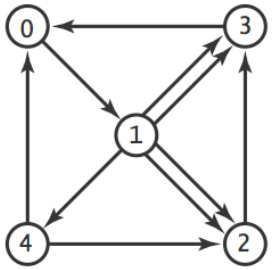



```
% more tiny.txt
5                  ← N
0 1
1 2   1 2
1 3   1 3   1 4
2 3
3 0                      ↖ links
4 0   4 2
```

Transition Matrix

Transition matrix. `p[i][j]` = prob. that surfer moves from page `i` to `j`.

input graph




```
5
0 1
1 2   1 2
1 3   1 3   1 4
2 3
3 0
4 0   4 2
```

link counts

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$

degrees

$$\begin{bmatrix} 1 \\ 5 \\ 1 \\ 1 \\ 2 \end{bmatrix}$$

leap probabilities

$$\begin{bmatrix} .02 & .02 & .02 & .02 & .02 \\ .02 & .02 & .02 & .02 & .02 \\ .02 & .02 & .02 & .02 & .02 \\ .02 & .02 & .02 & .02 & .02 \\ .02 & .02 & .02 & .02 & .02 \end{bmatrix} +$$

link probabilities

$$\begin{bmatrix} 0 & .90 & 0 & 0 & 0 \\ 0 & 0 & .36 & .36 & .18 \\ 0 & 0 & 0 & .90 & 0 \\ .90 & 0 & 0 & 0 & 0 \\ .45 & 0 & .45 & 0 & 0 \end{bmatrix} =$$

transition matrix

$$\begin{bmatrix} .02 & .92 & .02 & .02 & .02 \\ .02 & .02 & .38 & .38 & .20 \\ .02 & .02 & .02 & .92 & .02 \\ .92 & .02 & .02 & .02 & .02 \\ .47 & .02 & .47 & .02 & .02 \end{bmatrix}$$

surfer on page 1 goes to page 2 next 38% of the time

Monte Carlo Simulation

Monte Carlo simulation.

How? see next slide

- Surfer starts on page 0.
- Repeatedly choose next page, according to transition matrix.
- Calculate how often surfer visits each page.

page

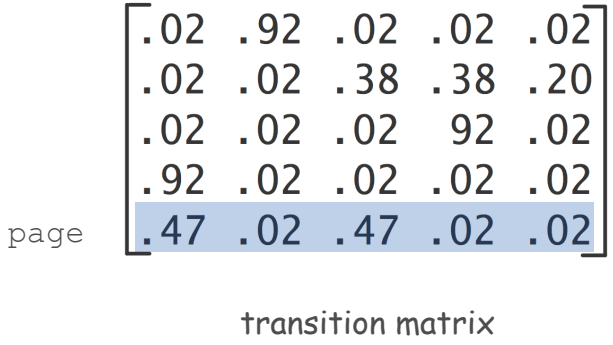
$$\begin{bmatrix} .02 & .92 & .02 & .02 & .02 \\ .02 & .02 & .38 & .38 & .20 \\ .02 & .02 & .02 & .92 & .02 \\ .92 & .02 & .02 & .02 & .02 \\ .47 & .02 & .47 & .02 & .02 \end{bmatrix}$$

transition matrix

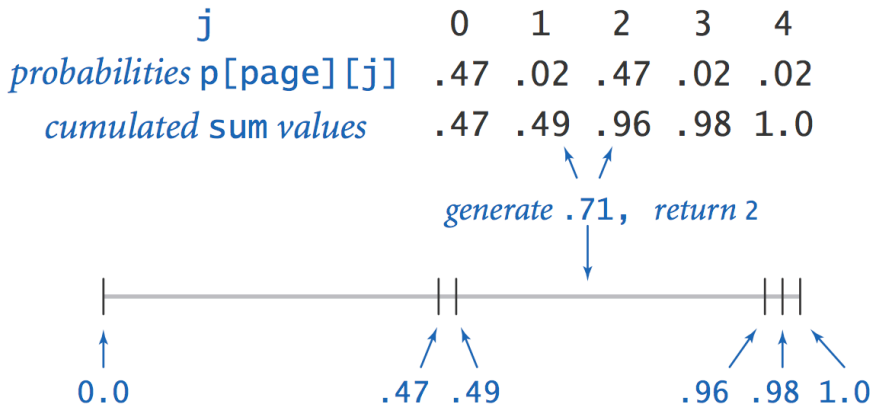
Random Surfer

Random move. Surfer is on page `page`. How to choose next page `j`?

- Row `page` of transition matrix gives probabilities.
- Compute cumulative probabilities for row `page`.
- Generate random number `r` between `0.0` and `1.0`.
- Choose page `j` corresponding to interval where `r` lies.



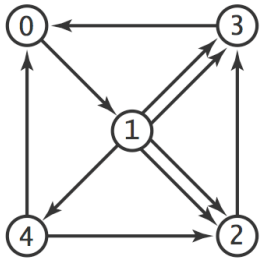
transition matrix



Mathematical Context

Convergence. For the random surfer model, the fraction of time the surfer spends on each page converges to a unique distribution, independent of the starting page.

"page rank"
"stationary distribution" of Markov chain
"principal eigenvector" of transition matrix



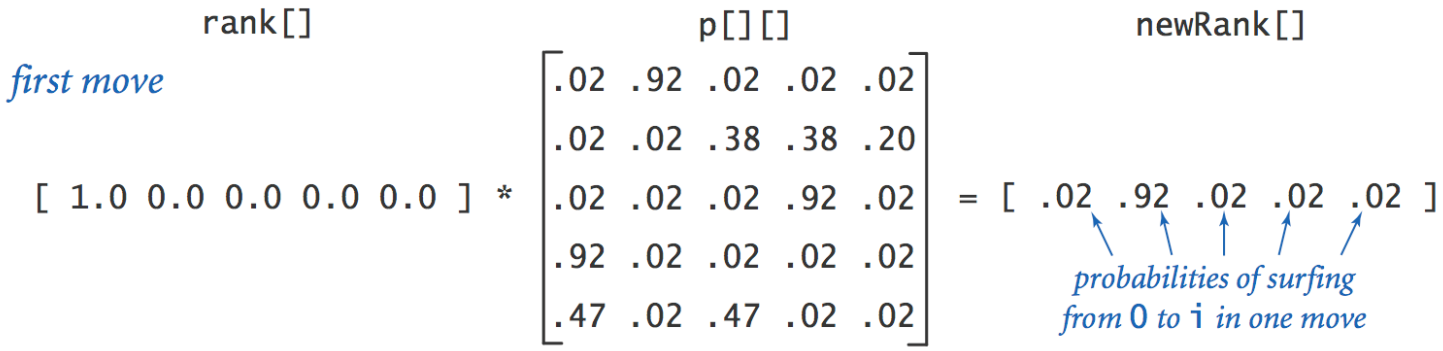
0 .27
1 .27
3 .25
2 .15
4 .07

$$\left[\frac{428{,}671}{1{,}570{,}055}, \frac{417{,}205}{1{,}570{,}055}, \frac{229{,}519}{1{,}570{,}055}, \frac{388{,}162}{1{,}570{,}055}, \frac{106{,}498}{1{,}570{,}055}\right]$$

The Power Method

Q. If the surfer starts on page `0`, what is the probability that surfer ends up on page `i` after one step?

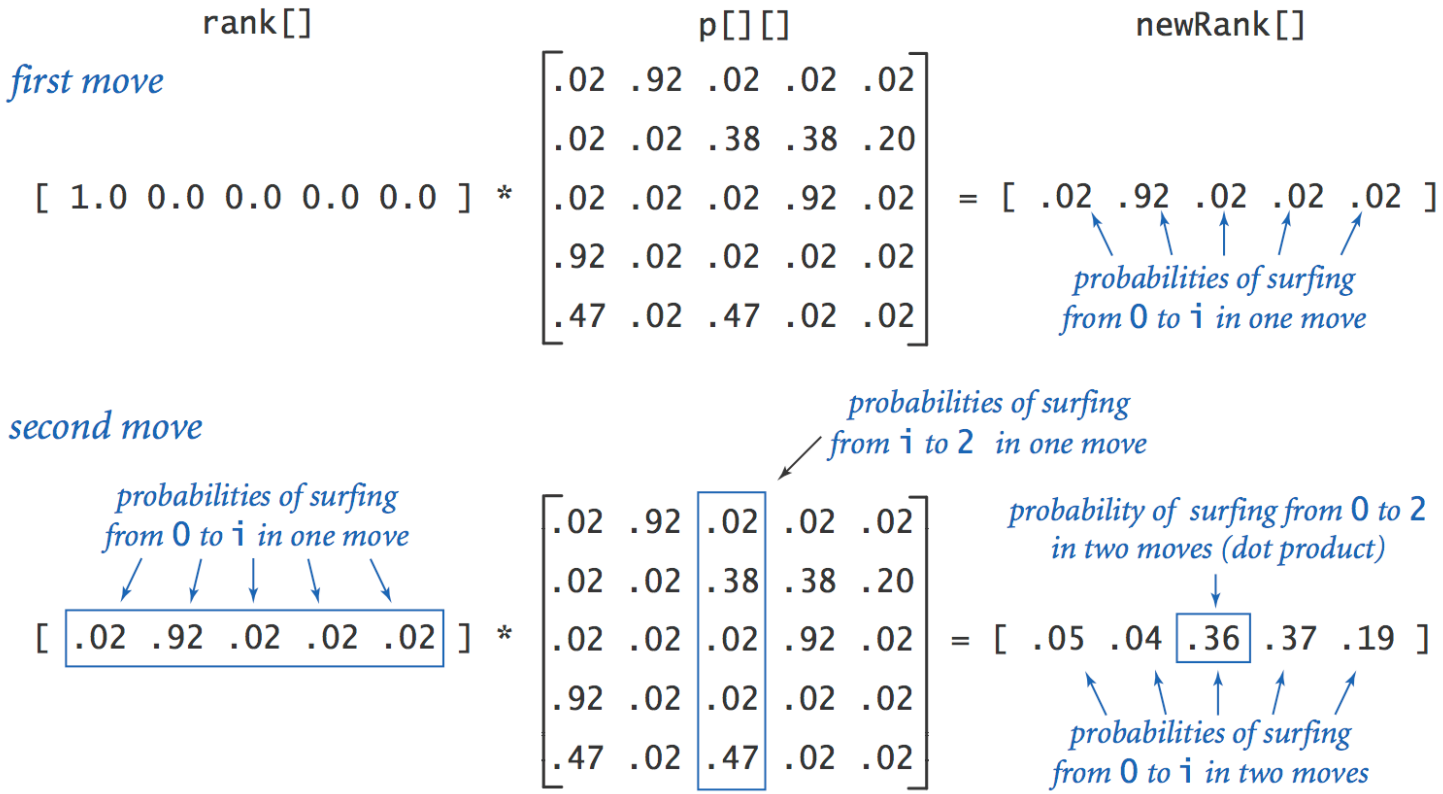
A. First row of transition matrix.



The Power Method

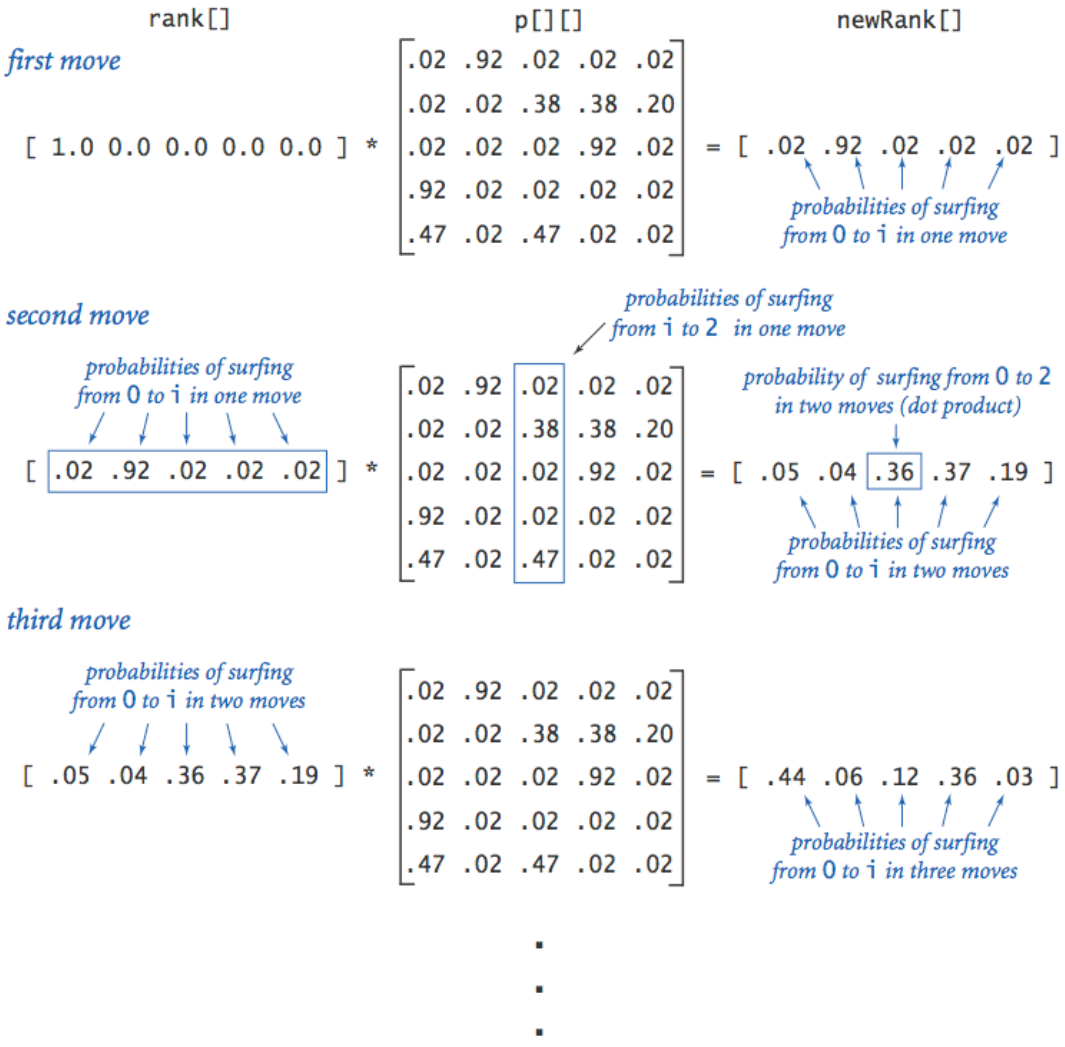
Q. If the surfer starts on page `0`, what is the probability that surfer ends up on page `i` after two steps?

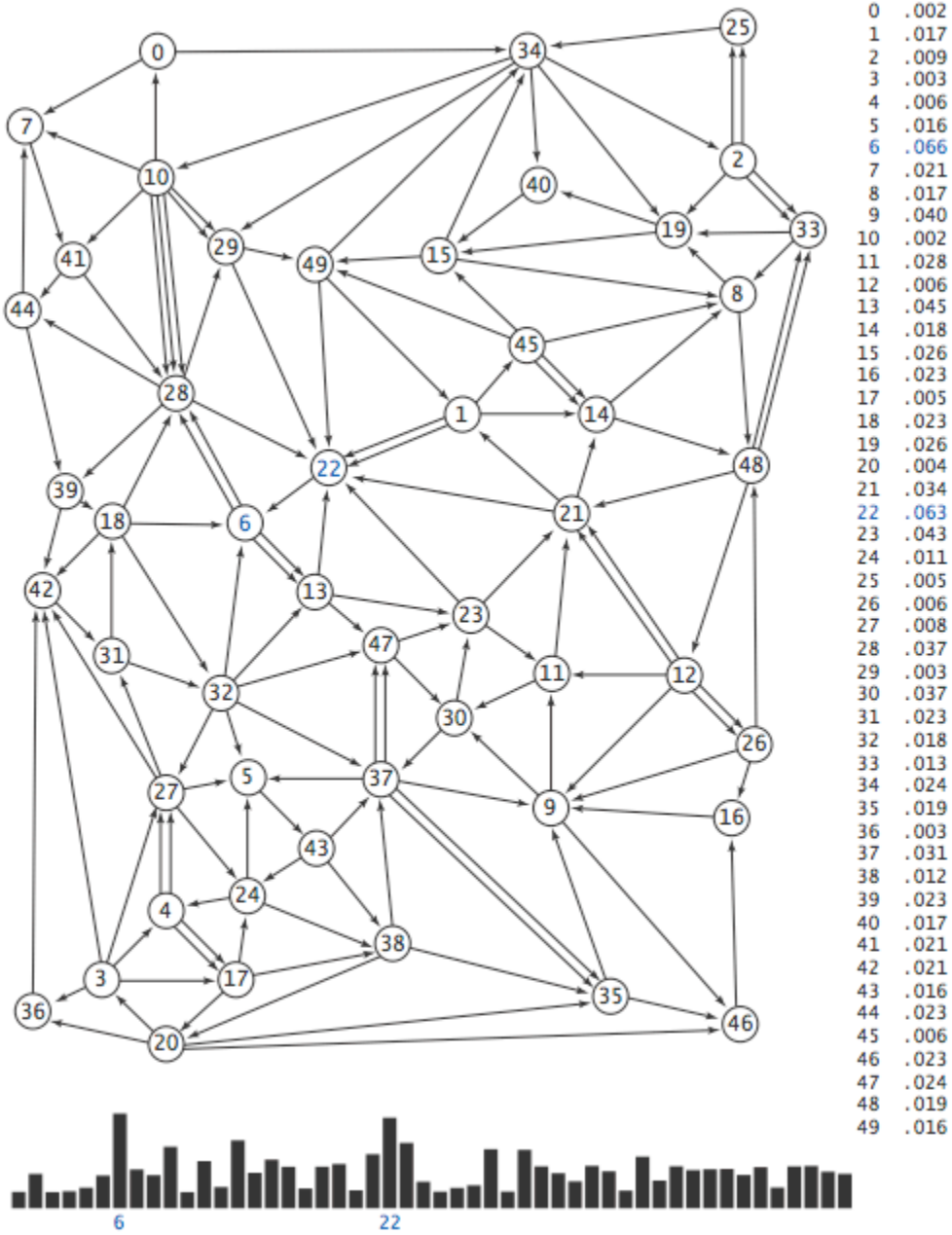
A. Matrix-vector multiplication.



The Power Method

Power method. Repeat until page ranks converge.



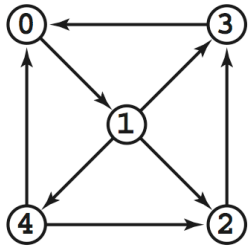


Page ranks with histogram for a larger example

Random Surfer: Scientific Challenges

Google's PageRank™ algorithm. [Sergey Brin and Larry Page, 1998]

- Rank importance of pages based on hyperlink structure of web, using 90-10 rule.
- Revolutionized access to world's information.



0 .269
1 .262
3 .227
2 .143
4 .099

Page ranks

Scientific challenges. Cope with 4 billion-by-4 billion matrix!

- Need data structures to enable computation.
- Need linear algebra to fully understand computation.