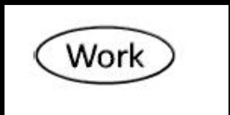
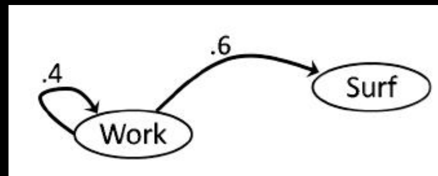
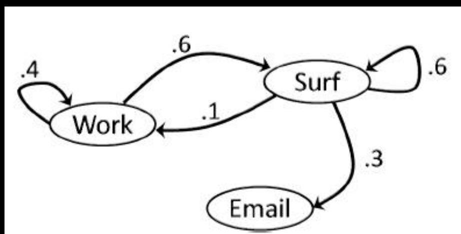
Markov Chains and PageRank

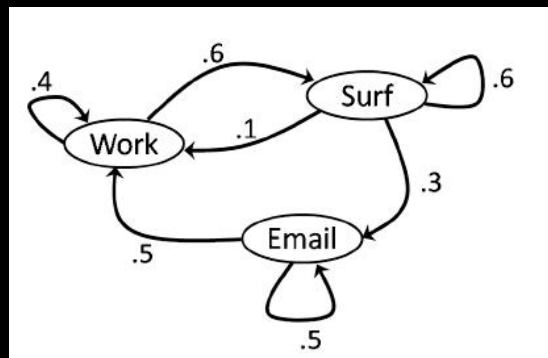
Anna Karlin

A first Markov chain



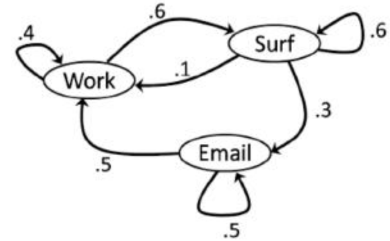






My daily life in a nutshell!

$$
\begin{array}{c c}
 & \begin{array}{ccc} W & S & E \end{array} \\
\begin{array}{c} W \\ S \\ E \end{array} &
\begin{pmatrix} .4 & .6 & 0 \\ .1 & .6 & .3 \\ .5 & 0 & .5 \end{pmatrix}
\end{array}
$$



Work
Surf
Email
.4
.6
.1
.6
.3
.5
.5

$$\begin{pmatrix} .4 & .6 & 0 \\ .1 & .6 & .3 \\ .5 & 0 & .5 \end{pmatrix}$$

$$P^2 = \begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .22 & .6 & .18 \\ .25 & .42 & .33 \\ .45 & .3 & .25 \end{pmatrix} \end{array}$$

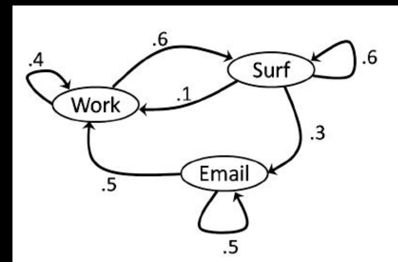
$$P^3 = \begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .238 & .492 & .270 \\ .307 & .402 & .291 \\ .335 & .450 & .215 \end{pmatrix} \end{array}$$

$$P^{10} \approx \begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .2940 & .4413 & .2648 \\ .2942 & .4411 & .2648 \\ .2942 & .4413 & .2648 \end{pmatrix} \end{array}$$

$$P^{30} \approx \begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .29411764705 & .44117647059 & .26470588235 \\ .29411764706 & .44117647058 & .26470588235 \\ .29411764706 & .44117647059 & .26470588235 \end{pmatrix} \end{array}$$

$$P^{60} \approx \begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .294117647058823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \end{pmatrix} \end{array}$$

$$\pi[S] = \frac{15}{34}, \pi[W] = \frac{10}{34}, \pi[E] = \frac{9}{34}.$$



$$\begin{cases} \pi[W] & = & .4\pi[W] + .1\pi[S] + .5\pi[E] \\ \pi[S] & = & .6\pi[W] + .6\pi[S] + \ 0\pi[E] \\ \pi[E] & = & \ 0\pi[W] + .3\pi[S] + .5\pi[E] \end{cases}$$

$$\pi[W] + \pi[S] + \pi[E] = 1$$

Fundamental Theorem of Markov Chains:

$\forall v$ long run probability of being in state v converges to $\pi[v]$

$$\pi[v] = \sum_u \pi[u] p_{uv}$$

Google and PageRank

from notes by Ryan O'Donnell

- 1997
  - Bill Clinton in White House
  - Deep Blue beat world chess champion (Kasparov)
  - And the Internet kind of sucked
  - Nov '97: only one of the top 4 commercial search engines actually *found itself* when you searched for it!

The Problem

- Search engines worked by matching words
- Top search for Bill Clinton
  - `Bill Clinton Joke of the Day' Website
- Deeply susceptible to spammers and advertisers

How to fix?

- Collect pages with decent textual match
- Then **rank** them by some measure of ‘quality’ or ‘authority’.

- Enter two groups:
  - Jon Kleinberg (prof at Cornell)
  - Larry Page and Sergey Brin (Ph.D. students at Stanford)

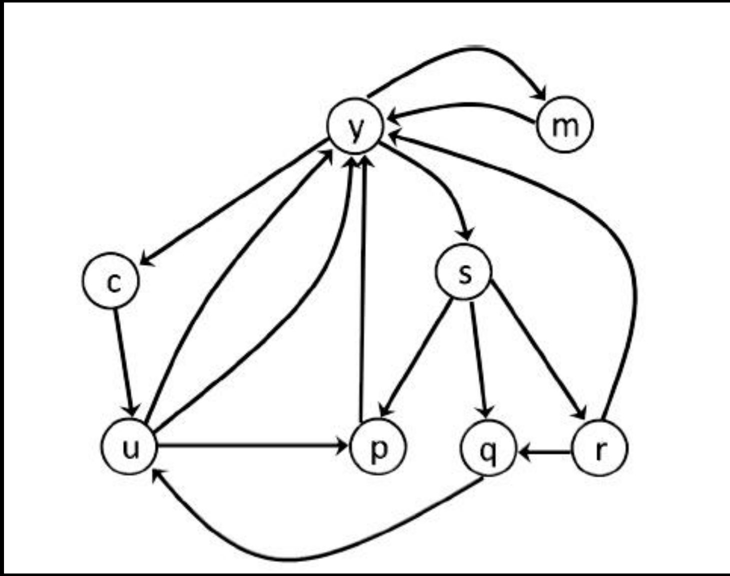
Both had pretty much same brilliant idea … and it worked!

Two groups:

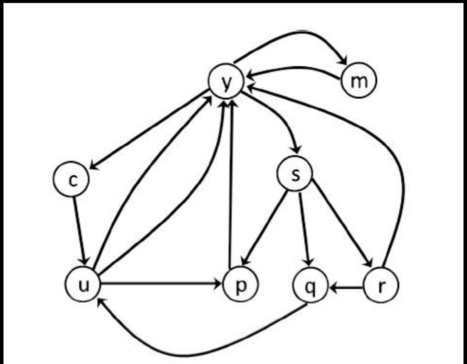
- Larry Page and Sergey Brin (Ph.D. students at Stanford)
  - Took the idea and founded Google, making billions.
- Jon Kleinberg (prof at Cornell)
  - MacArthur Genius Prize, Nevanlinna Prize, many academic honors

PageRank

- Key idea: hyperlink analysis: take into account directed graph structure of the web.



PageRank



- Idea1: "Citations"
  - As with academic publishing, it's a good idea to think of each link to a page as a "citation" or "vote of quality".
  - Rank pages by in-degree?

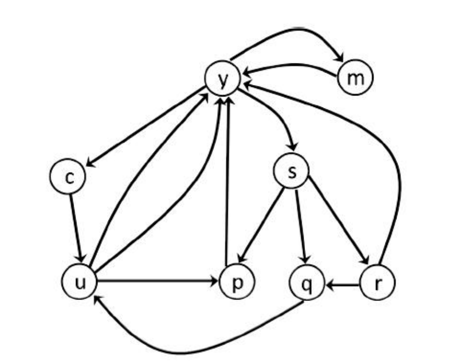
Rank pages by in-degree

- Problem:
  - Spamming
  - Some linkers are not terribly discriminating
  - Not all links are created equal.

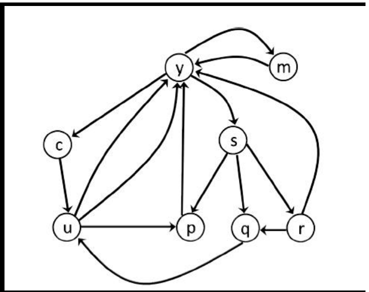
- Perhaps we should weight the links somehow and then use the weights of the in-links to rank pages.

Idea that works well

- Web page has high quality if it's linked to by lots of high quality pages.
- A page is high quality if it links to lots of high quality pages.
- So kind of a recursive definition



Page and Brin's Idea

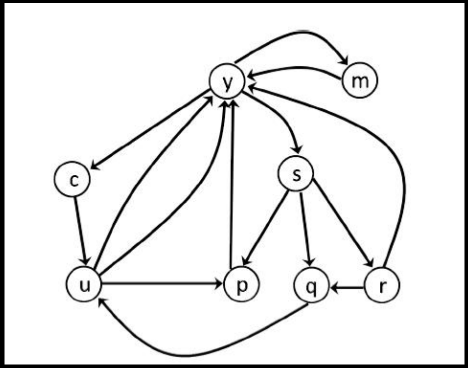


- Links convey authority.
- The linker themselves is authoritative if they are heavily linked to (i.e. well-connected in graph)
- Idea: Imagine a "random web surfer"
  - At each step looking at some web page.
  - Transitions to next web page by following a random link from that page.
  - Authority/rank of page: long run probability of being at that page = stationary probability

Fundamental Theorem of Markov Chains:

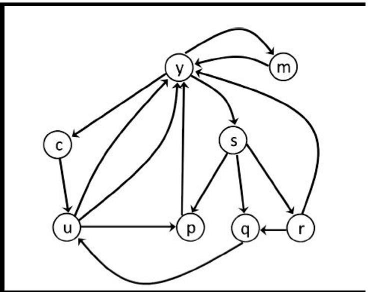
$\forall v$ long run probability of being in state v converges to $\pi[v]$

$$\pi[v] = \sum_u \pi[u] p_{uv}$$



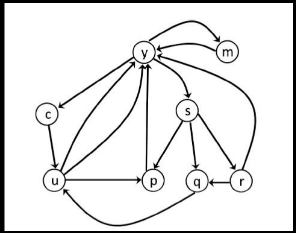
Random surfer!!

Page and Brin's Idea



- Idea: Imagine a "random web surfer"
  - Compute stationary probabilities for all pages on the web.
  - On a query, find pages containing the query terms. Return those pages, ranked in decreasing order of stationary probability.

Problems with Markov Chain approach



- Web pages with no outlinks.
- Spamming: add an entire group of nodes that all link only to each other.

Final Model

- Random surfer model (random walk):
  - On each step, with probability p follow a random real link on the current page
  - With probability 1-p go to a completely random page in the entire web.

Solve this sytem:

$$\pi[v] = \sum_u \pi[u] p_{uv}$$

Recursive definition always has a unique solution: called stationary distribution of the Markov chain

This was the idea on which Google was founded

- Since 1997, lots more secret sauce added….