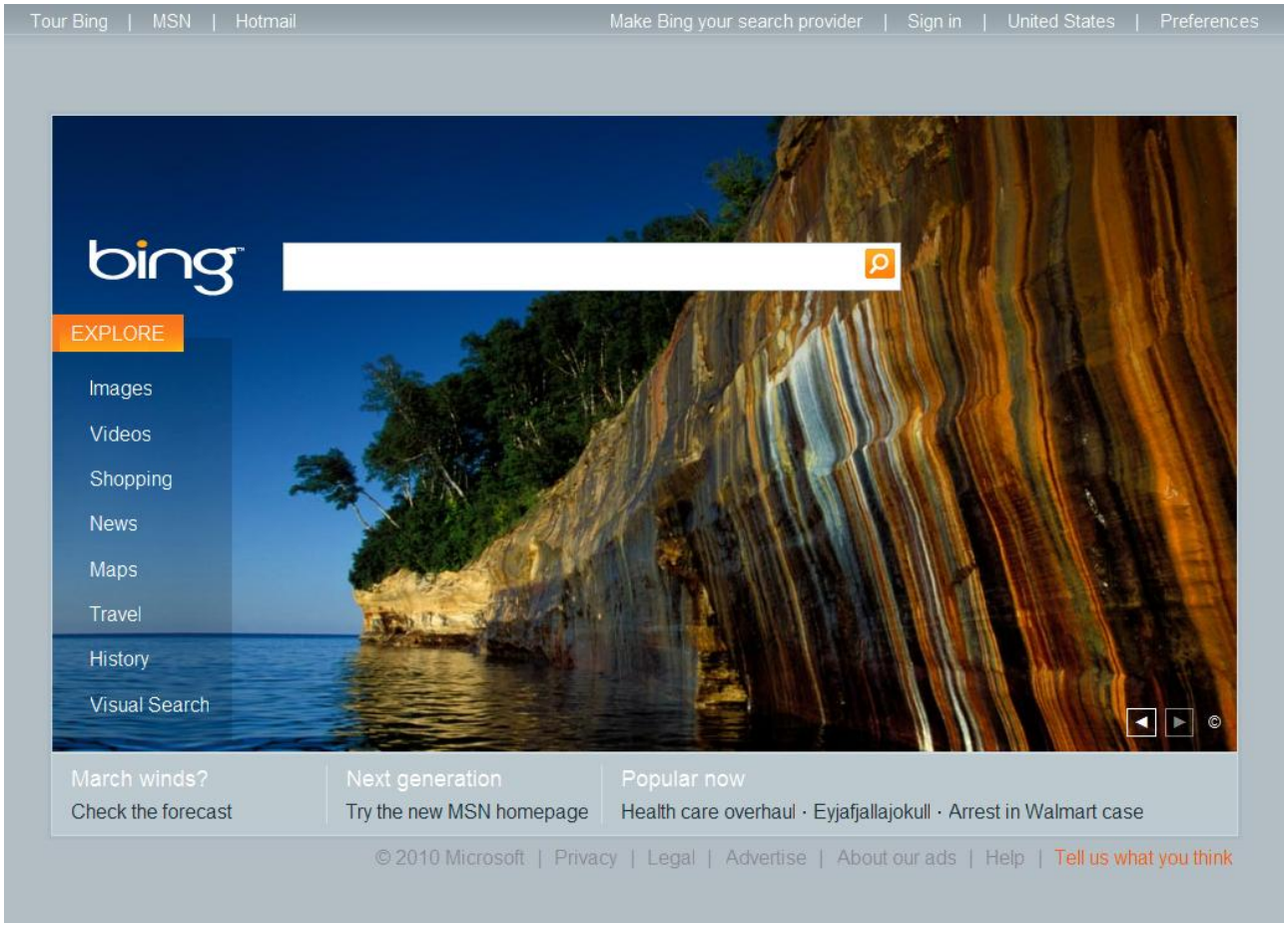
Mathematics of the Web

Prof. Sara Billey

University of Washington

Search Engines



Google™

YAHOO!

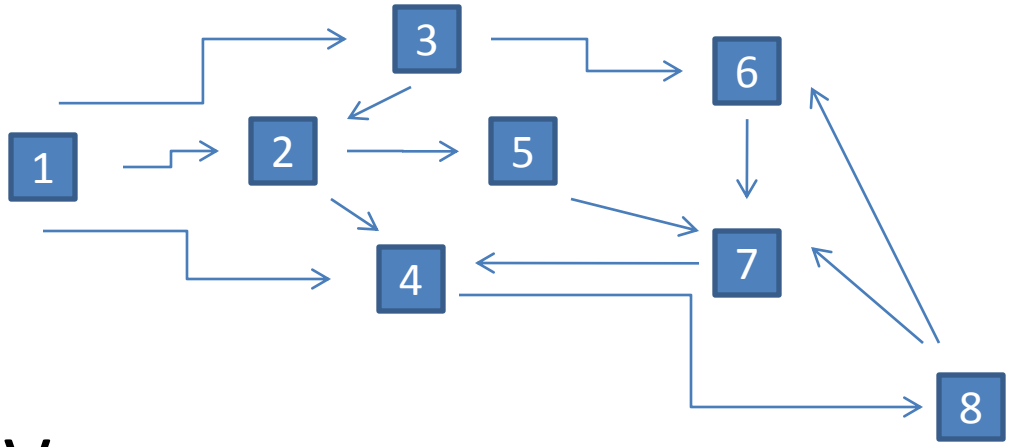
Real Problems/Mathematical Solutions

- When you type in a query, how does Google or Bing or Yahoo or any other search engine choose what to show you first?
- When there are many web sites mirroring the same data, how does the search engine know enough to only show you one of those pages?
- How does a search engine know which images correspond with your search query?
- How big is the web?

Graph Theory

The mathematical model of the web is a graph:

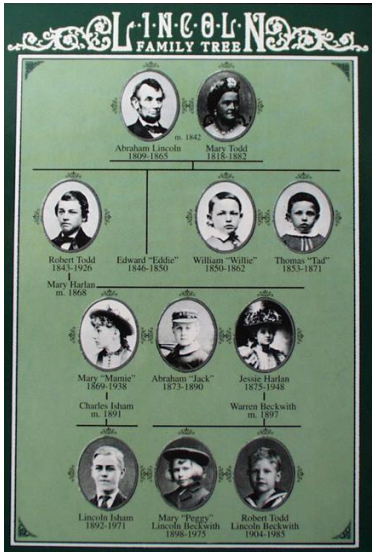
G= (Vertices , Edges)



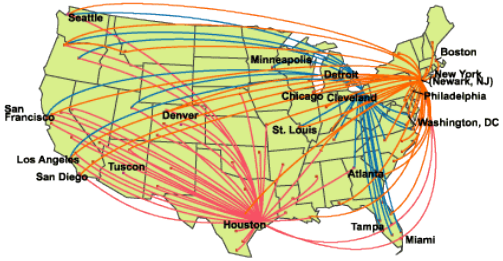
Vertices = V = {1,2,3,4,5,6,7,8}

Edges = E = arrows = {(1,2),(1,3),(1,4),(2,4), (2,5), (3,2), (3,6), (4, 8), (5,7),(6,7),(7,4),(8,6),(8,7)}

Real World Graphs



- Family Trees
- Chain of command in the military
- Airline routes



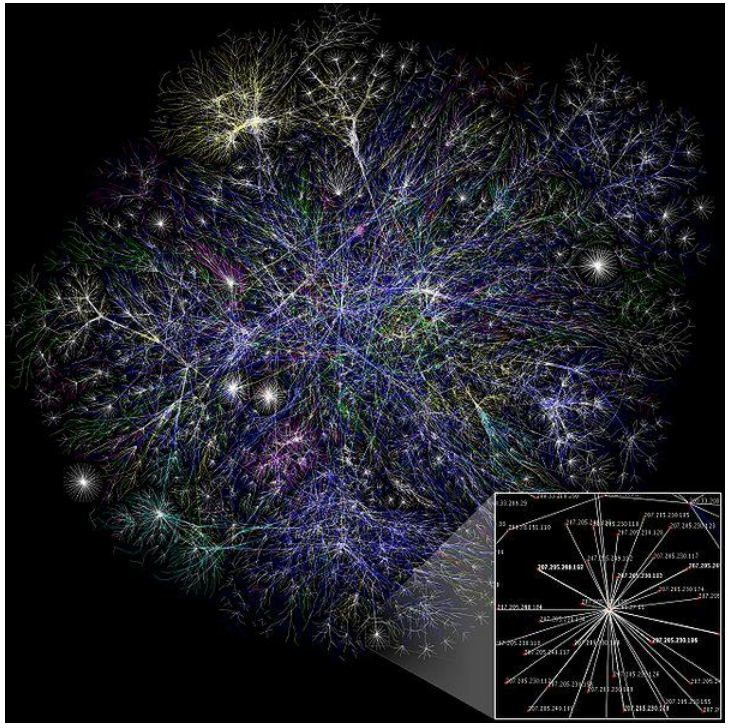
- Cell phones/contact info
- Facebook pages--

Graph Theory

Graphs encode relationships between pairs of objects in a set.

V = IP Addresses

E= Webpage links



Visualization of the various routes through a portion of the Internet [Wikipedia]

Question I

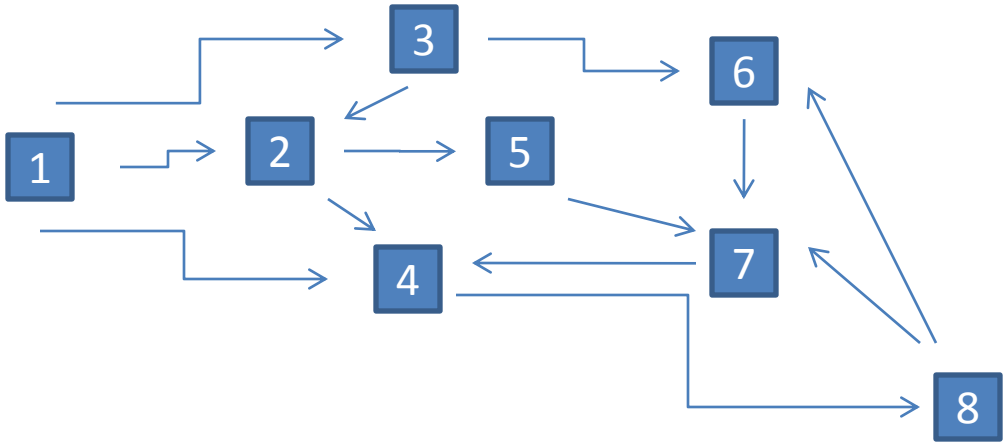
When you type in a query, how does the search engine choose what to show you first?

Answer:

- Quickly identify all web pages that contain your query.
- Decide which pages to show you first, second, third, …., top 10, top 100, etc .

Question I

When you type in a query, how does the search engine choose what to show you first?



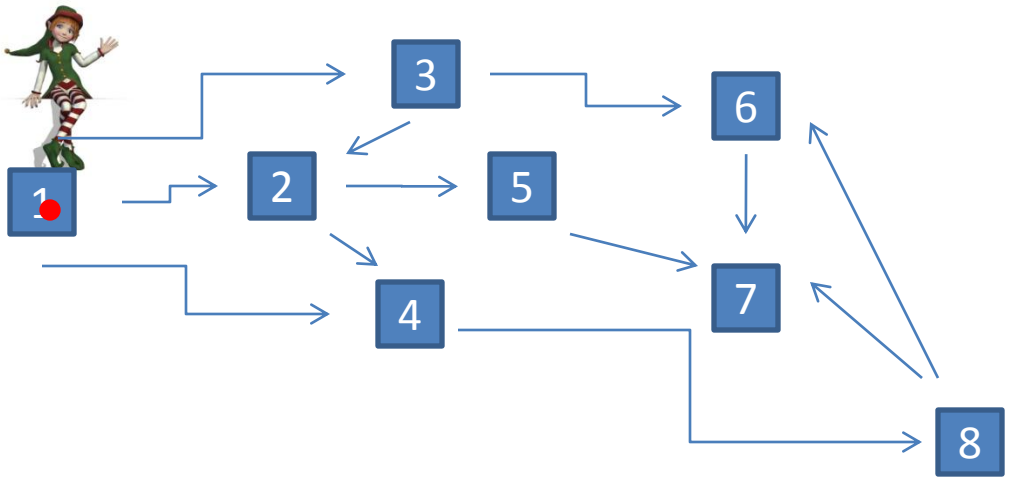
Does any vertex in this graph stand out as first to you?

Random Surfer

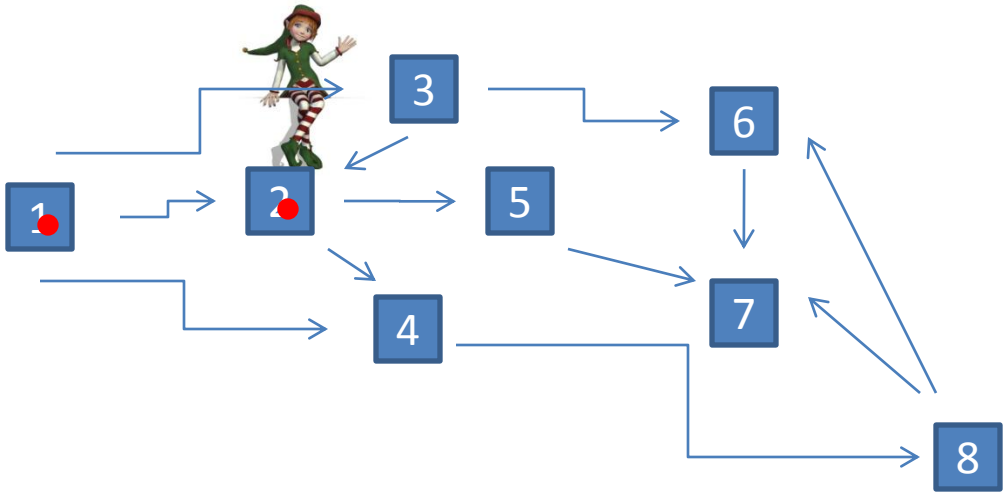
Experiment:

- Each of you start at your favorite website.
- Choose any link on that page at random and follow it to the next page. Choose any link on that page and follow it, etc.
- Record how often you visit each site.

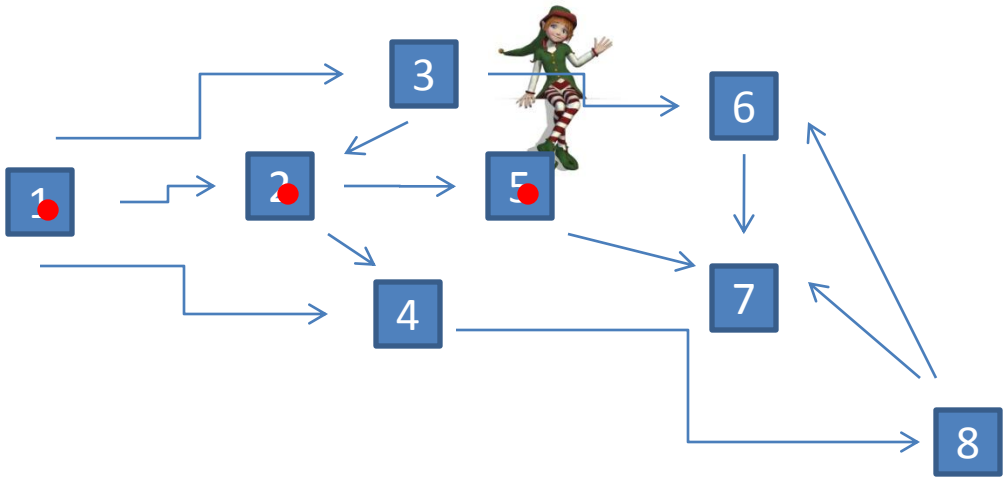
Random Surfer



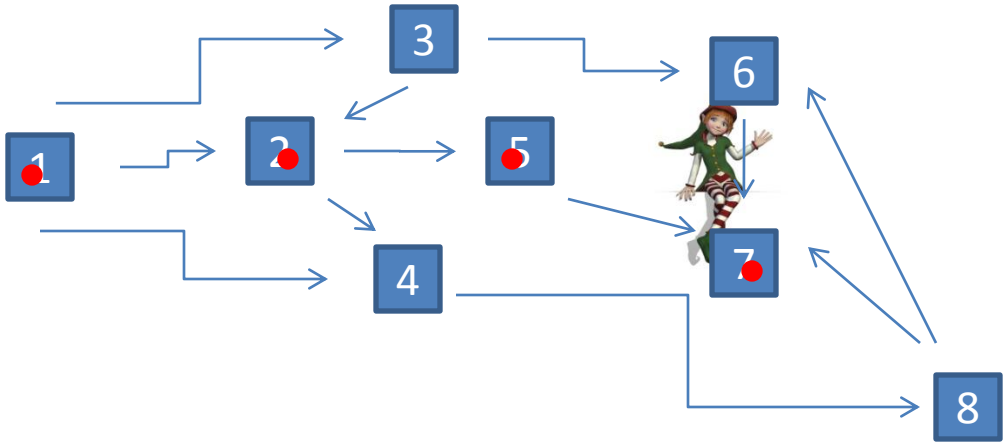
Random Surfer



Random Surfer



Random Surfer



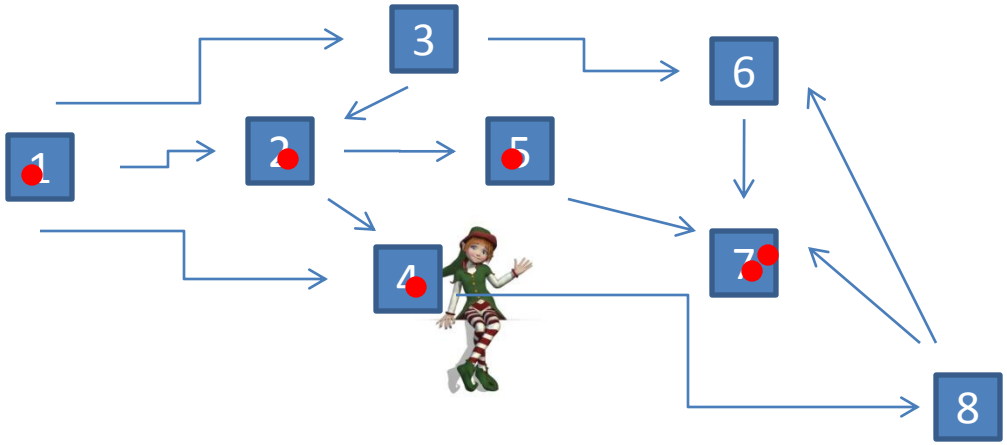
Oops! She is stuck.

Random Surfer

New Experiment:

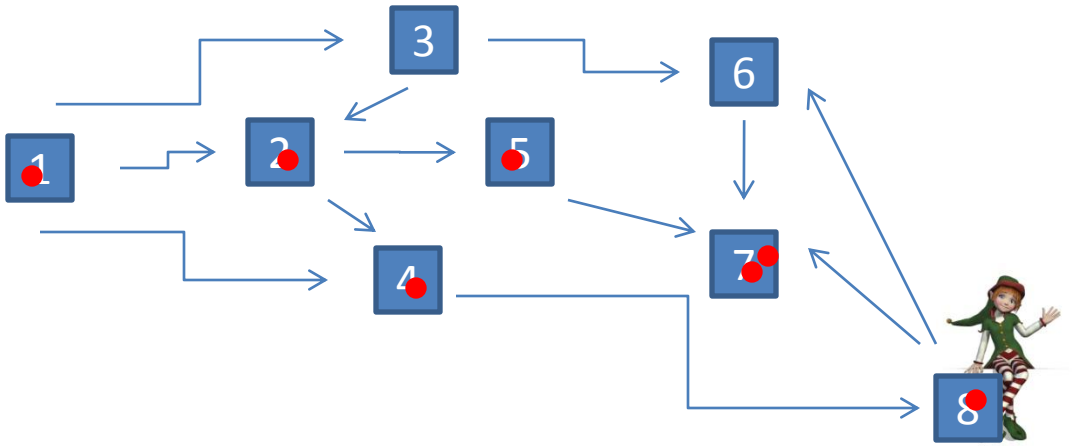
- Each of you start at your favorite website.
- With probability p=.15, jump to any website in the graph. Otherwise, with probability q=.85, choose any link on that page at random and follow it to the next page, including a link from each page to itself. Repeat!
- Record how often you visit each site.

Random Surfer

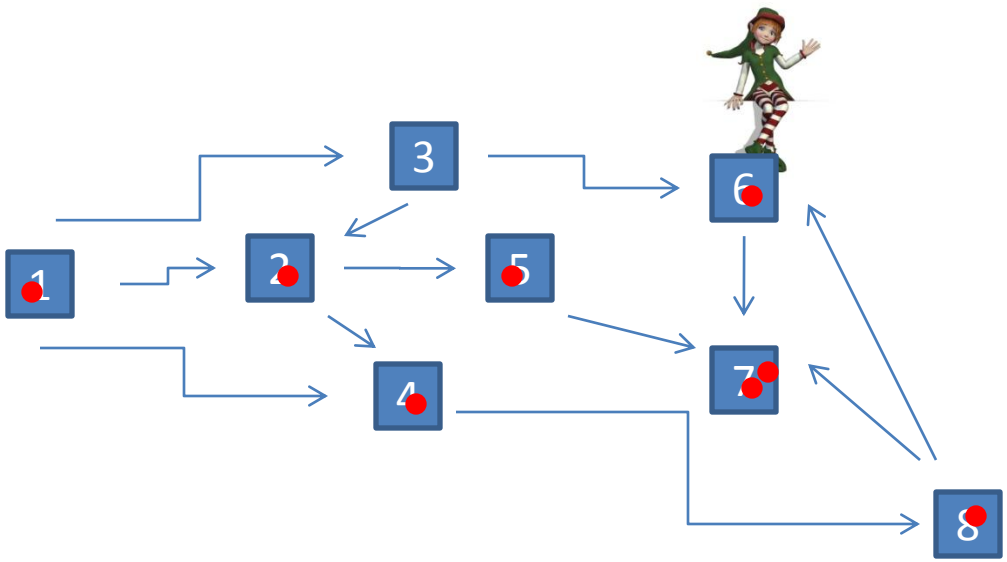


After a few turns, she jumps to a random web page: say 4.

Random Surfer



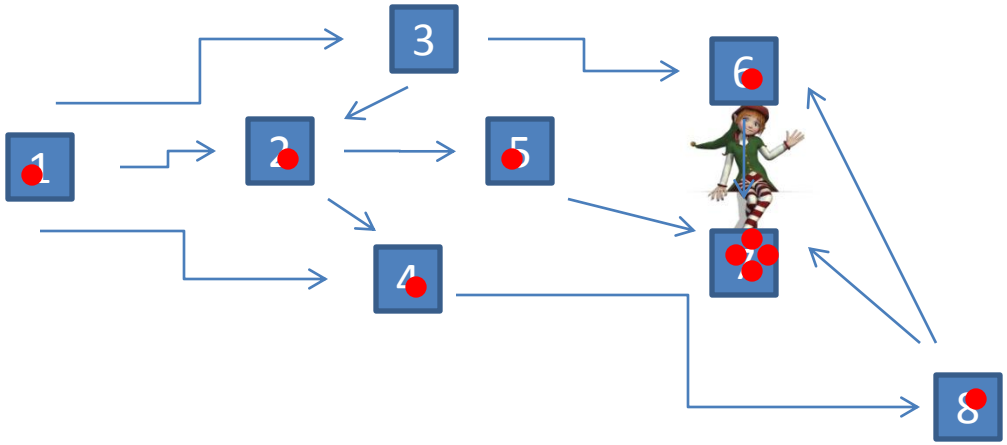
Random Surfer



Random Surfer

After 10 steps…

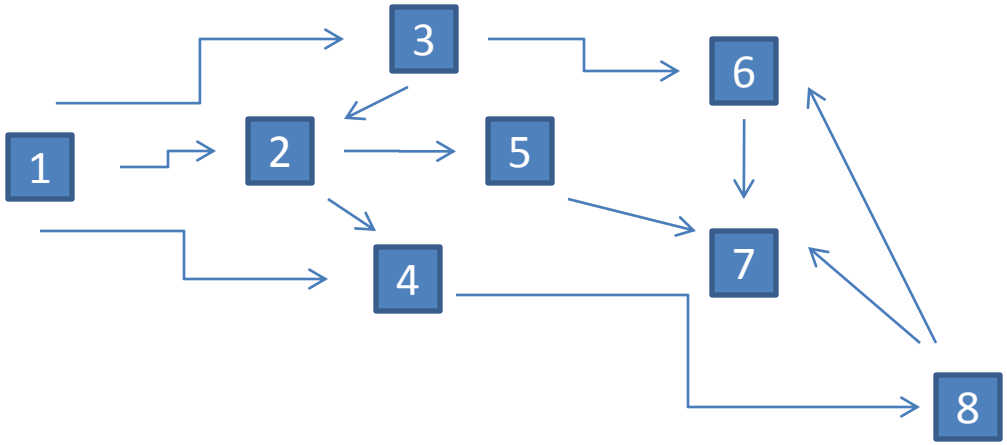
Page	Count
1	1
2	1
3	0
4	1
5	1
6	1
7	4
8	1



Random Surfer

After 100 steps…

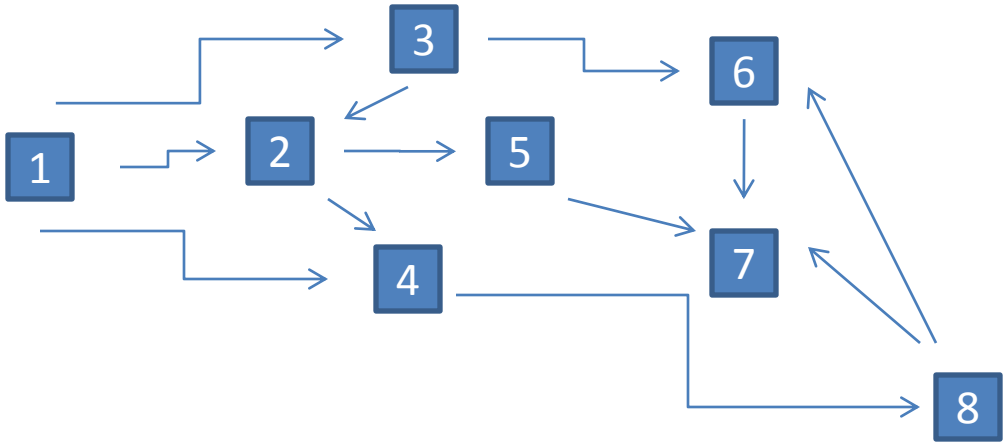
Page	Count
1	5
2	3
3	2
4	4
5	9
6	6
7	69
8	2



Random Surfer

After 1000 steps…

Page	Count
1	24
2	55
3	26
4	67
5	43
6	65
7	651
8	69



Convergence

After 10 steps…

Page	Count
1	1
2	1
3	0
4	1
5	1
6	1
7	4
8	1

After 100 steps…

Page	Count
1	5
2	4
3	1
4	8
5	2
6	2
7	67
8	11

After 1000 steps…

Page	Count
1	24
2	55
3	26
4	67
5	43
6	65
7	651
8	69

After 10,000 steps…

Page	Count
1	258
2	443
3	314
4	624
5	529
6	826
7	6341
8	665

Convergence

Given enough iterations we see that the frequency of visiting each site stabilizes!

Page	Count
1	258
2	443
3	314
4	624
5	529
6	826
7	6341
8	665

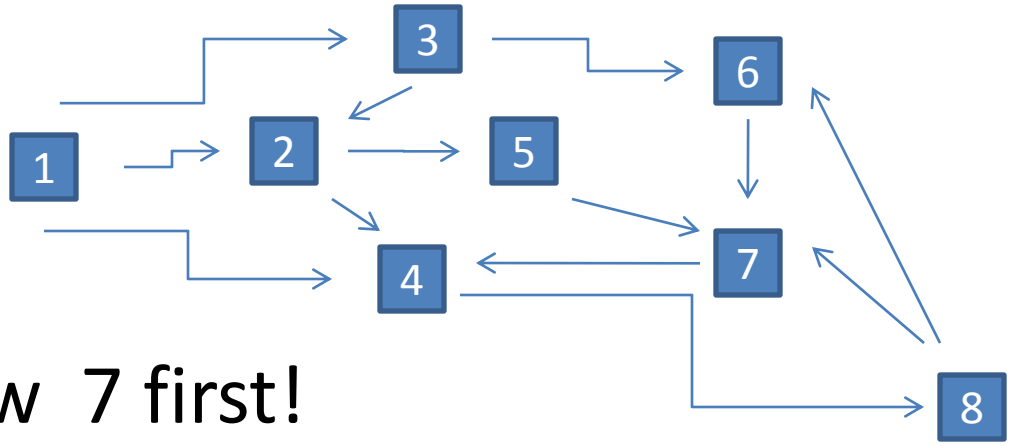
Page	%
1	2.58
2	4.43
3	3.14
4	6.24
5	5.29
6	8.26
7	63.41
8	6.65

Ah ha!
We can use these frequencies to rank web pages!

Question I

When you type in a query, how does the search engine choose what to show you first?

Page	%
1	2.58
2	4.43
3	3.14
4	6.24
5	5.29
6	8.26
7	63.41
8	6.65



Show 7 first!

Predicting Frequencies

How can we predict the frequency/probability that our random surfer will visit each page quickly?

Answer: Use the theory of Markov processes and Linear Algebra!

The Trick!

Let $pr_t(i)$ be the probability that our random surfer is at page i at time t for i=1,2,3,..,N=8.

Then iterating the following function approximates the probability distribution in the random surfer model:

$$pr_{t+1}(i) = \frac{.15}{N} + .85 \sum_{j->i} \frac{pr_t(j)}{\#links(j)}$$

Experimental vs Theoretical

Compare the percentage of visits our random surfer was on page i with the function $pr_{10}(i)$:

Page	%
1	2.58
2	4.43
3	3.14
4	6.24
5	5.29
6	8.26
7	63.41
8	6.65

Page	$Pr_{10}(i)$
1	.024
2	.046
3	.033
4	.064
5	.056
6	.082
7	.630
8	.065

Summary of PageRank

When you type in a query, how does a search engine choose what to show you first?

- One approach: Select all web pages containing the query and show pages in order so $pr_t(i)$ is decreasing!
- Linear algebra, probability and Markov processes make this algorithm efficient even on billions of web pages.
- This algorithm invented by Larry Page and Sergey Brin [1998], founders of Google.
- Other approaches exist: HITS by Jon Kleinberg

Question II

When there are many web sites mirroring the same data, how does the search engine know enough to only show you one of those pages?

Answer: Compare Fingerprints!

Are these the same?

Mr. and Mrs. Dursley of number four, Privet Drive, were proud to say, they were perfectly normal, thank you very much. They were the last people you would expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

…

Mr. and Mrs. Dursley of number five, Privet Drive, were proud to say, they were perfectly normal, thank you very much. They were the first people you would expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

…

Fingerprinting Web Pages

1. Shingle each document: consider the first N strings of K words in each document.

Example shingles (K=4):

- Mr. and Mrs. Dursley
- and Mrs. Dursley of
- Mrs. Dursley of number
- Dursley of number four

Fingerprinting Web Pages

1. Shingle each document: consider the first N strings of K words in each document.
2. List all such strings of K words that occur in the entire indexable web in order: $[s_1, s_2, \ldots s_z]$.
3. If a documents starts out: $S_{1540}, S_{298}, S_{1730}$, remember only 1540, 298, 1730.
4. Choose a random permutation of 1,2,…, Z p:{1,2,…, Z} → {1,2,…, Z}.
5. For each document, remember only the minimum {p(1540),p(298),p(1730)}.

Real Problems/Mathematical Solutions

- When you type in a query, how does Google or Bing or Yahoo or any other search engine choose what to show you first? Pagerank , HITS
- When there are many web sites mirroring the same data, how does the search engine know enough to only show you one of those pages? Broder Algorithm
- How does a search engine know which images correspond with your search query? Photosynth, Snavely-Simon-Goesele-Szeliski-Seitz
- How big is the web? Gulli and Signorini

The End

To find a copy of these notes on the web, just search for "Sara math" or "Billey"