

CS6200  
Information Retrieval

**PageRank Continued**

*with slides from  
Hinrich Schütze and Christina Lioma*

# Exercise: Assumptions underlying PageRank

- Assumption 1: A link on the web is a quality signal – the author of the link thinks that the linked-to page is high-quality.
- Assumption 2: The anchor text describes the content of the linked-to page.
- Is assumption 1 true in general?
- Is assumption 2 true in general?

# Google bombs

# Google bombs

- A Google bomb is a search with “bad” results due to maliciously manipulated anchor text.

# Google bombs

- A Google bomb is a search with “bad” results due to maliciously manipulated anchor text.
- Google introduced a new weighting function in January 2007 that fixed many Google bombs.

# Google bombs

- A Google bomb is a search with “bad” results due to maliciously manipulated anchor text.
- Google introduced a new weighting function in January 2007 that fixed many Google bombs.
- Still some remnants: [dangerous cult] on Google, Bing, Yahoo

# Google bombs

- A Google bomb is a search with “bad” results due to maliciously manipulated anchor text.
- Google introduced a new weighting function in January 2007 that fixed many Google bombs.
- Still some remnants: [dangerous cult] on Google, Bing, Yahoo
  - Coordinated link creation by those who dislike the Church of Scientology

# Google bombs

- A Google bomb is a search with “bad” results due to maliciously manipulated anchor text.
- Google introduced a new weighting function in January 2007 that fixed many Google bombs.
- Still some remnants: [dangerous cult] on Google, Bing, Yahoo
  - Coordinated link creation by those who dislike the Church of Scientology
- Defused Google bombs: [dumb motherf...], [who is a failure?], [evil empire]



# Origins of PageRank: Citation analysis (1)

- Citation analysis: analysis of citations in the scientific literature.
- Example citation: “[Miller \(2001\)](#) has shown that physical activity alters the metabolism of estrogens.”
- We can view “Miller (2001)” as a hyperlink linking two scientific articles.
- One application of these “hyperlinks” in the scientific literature:
  - Measure the similarity of two articles by the overlap of other articles citing them.
  - This is called [cocitation similarity](#).
  - Cocitation similarity on the web: Google’s “find pages like this” or “Similar” feature.

# Origins of PageRank: Citation analysis (2)

- Another application: Citation frequency can be used to measure the **impact** of an article .
  - Simplest measure: Each article gets one vote – not very accurate.
- On the web: citation frequency = **inlink count**
  - A high inlink count does not necessarily mean high quality ...
  - ... mainly because of link spam.
- Better measure: **weighted** citation frequency or citation rank
  - An article's vote is weighted according to its citation impact.
  - Circular? No: can be formalized in a well-defined way.

# Origins of PageRank: Citation analysis (3)

- Better measure: weighted citation frequency or citation rank.
- This is basically PageRank.
- PageRank was invented in the context of citation analysis by Pinski and Narin in the 1960s.
- Citation analysis is a big deal: The budget and salary of this lecturer are / will be determined by the impact of his publications!

# Origins of PageRank: Summary

- We can use the same formal representation for
  - citations in the scientific literature
  - hyperlinks on the web
- Appropriately weighted citation frequency is an excellent measure of quality ...
  - ... both for web pages and for scientific publications.
- Next: PageRank algorithm for computing weighted citation frequency on the web.

# Model behind PageRank: Random walk

- Imagine a web surfer doing a random walk on the web
  - Start at a random page
  - At each step, go out of the current page along one of the links on that page, equiprobably
- In the steady state, each page has a **long-term visit rate**.
- This long-term visit rate is the page's **PageRank**.
- **PageRank = long-term visit rate = steady state probability.**

# Formalization of random walk: Markov chains

# Formalization of random walk: Markov chains

- A Markov chain consists of  $N$  states, plus an  $N \times N$  transition probability matrix  $P$ .

# Formalization of random walk: Markov chains

- A Markov chain consists of  $N$  states, plus an  $N \times N$  transition probability matrix  $P$ .
- state = page

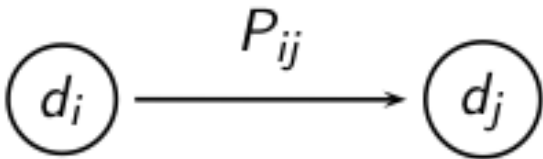


# Formalization of random walk: Markov chains

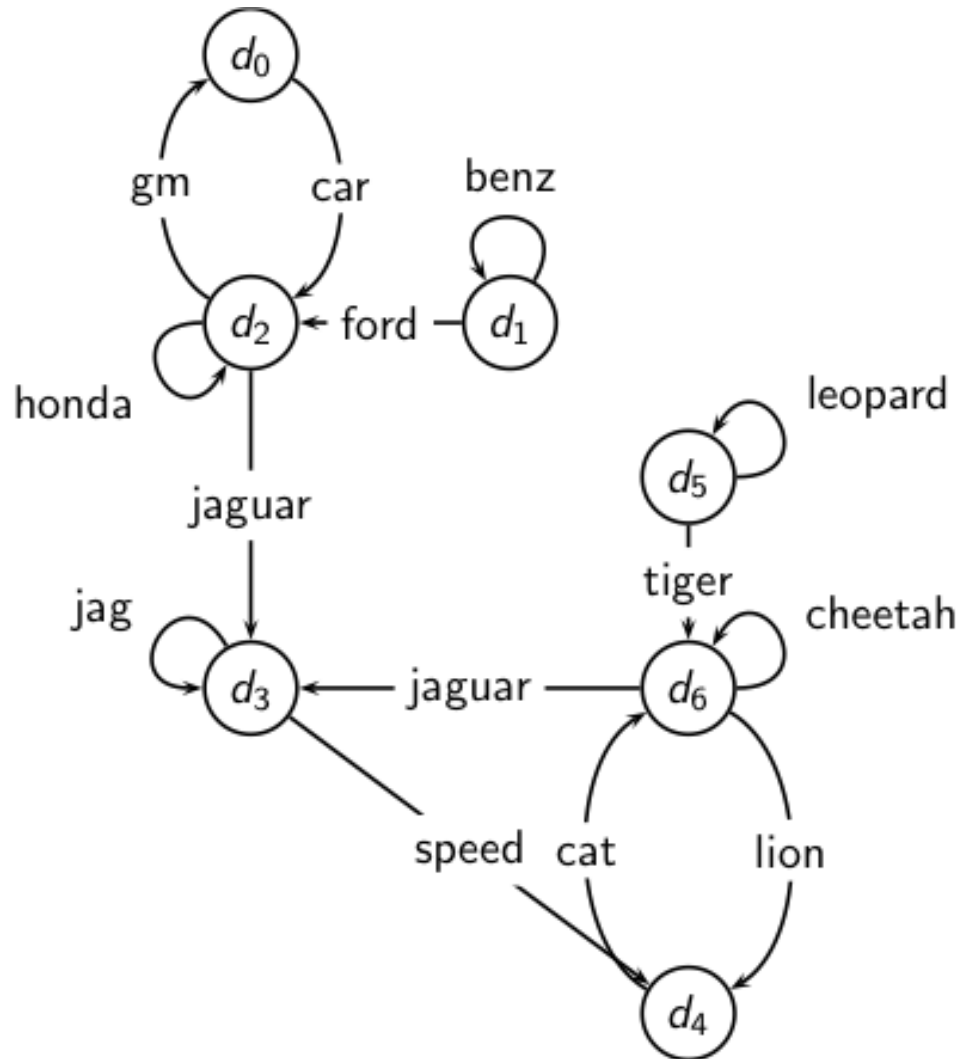
- A Markov chain consists of  $N$  states, plus an  $N \times N$  transition probability matrix  $P$ .
- state = page
- At each step, we are on exactly one of the pages.

# Formalization of random walk: Markov chains

- A Markov chain consists of  $N$  states, plus an  $N \times N$  transition probability matrix  $P$ .
- state = page
- At each step, we are on exactly one of the pages.
- For  $1 \leq i, j \leq N$ , the matrix entry  $P_{ij}$  tells us the probability of  $j$  being the next page, given we are currently on page  $i$ .
- Clearly, for all  $i$ ,  $\sum_{j=1}^N P_{ij} = 1$



# Example web graph



# Link matrix for example

# Link matrix for example

	$d_0$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
$d_0$	0	0	1	0	0	0	0
$d_1$	0	1	1	0	0	0	0
$d_2$	1	0	1	1	0	0	0
$d_3$	0	0	0	1	1	0	0
$d_4$	0	0	0	0	0	0	1
$d_5$	0	0	0	0	0	1	1
$d_6$	0	0	0	1	1	0	1

Transition probability matrix  $P$  for example

# Transition probability matrix $P$ for example

	$d_0$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$
$d_0$	0.00	0.00	1.00	0.00	0.00	0.00	0.00
$d_1$	0.00	0.50	0.50	0.00	0.00	0.00	0.00
$d_2$	0.33	0.00	0.33	0.33	0.00	0.00	0.00
$d_3$	0.00	0.00	0.00	0.50	0.50	0.00	0.00
$d_4$	0.00	0.00	0.00	0.00	0.00	0.00	1.00
$d_5$	0.00	0.00	0.00	0.00	0.00	0.50	0.50
$d_6$	0.00	0.00	0.00	0.33	0.33	0.00	0.33

# Long-term visit rate



# Long-term visit rate

- Recall: PageRank = long-term visit rate.

# Long-term visit rate

- Recall: PageRank = long-term visit rate.
- Long-term visit rate of page  $d$  is the probability that a web surfer is at page  $d$  at a given point in time.

# Long-term visit rate

- Recall: PageRank = long-term visit rate.
- Long-term visit rate of page  $d$  is the probability that a web surfer is at page  $d$  at a given point in time.
- Next: what properties must hold of the web graph for the long-term visit rate to be well defined?

# Long-term visit rate

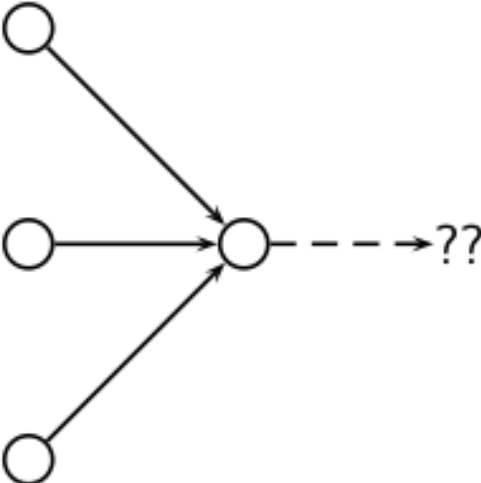
- Recall: PageRank = long-term visit rate.
- Long-term visit rate of page  $d$  is the probability that a web surfer is at page  $d$  at a given point in time.
- Next: what properties must hold of the web graph for the long-term visit rate to be well defined?
- The web graph must correspond to an **ergodic** Markov chain.

# Long-term visit rate

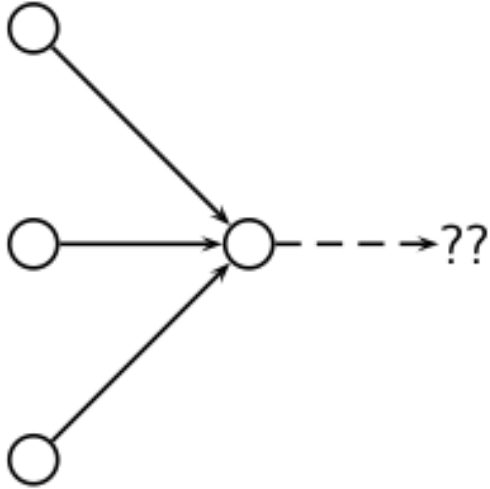
- Recall: PageRank = long-term visit rate.
- Long-term visit rate of page  $d$  is the probability that a web surfer is at page  $d$  at a given point in time.
- Next: what properties must hold of the web graph for the long-term visit rate to be well defined?
- The web graph must correspond to an **ergodic** Markov chain.
- First a special case: The web graph must not contain **dead ends**.

Dead ends

# Dead ends



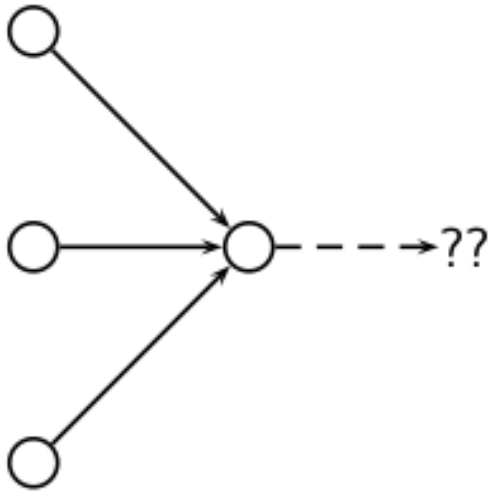
# Dead ends



- The web is full of dead ends.

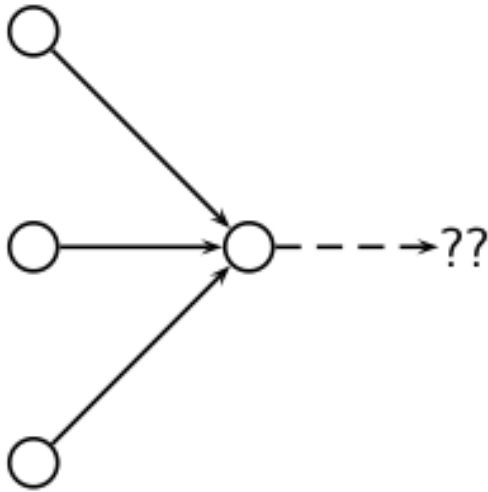


# Dead ends



- The web is full of dead ends.
- Random walk can get stuck in dead ends.

# Dead ends



- The web is full of dead ends.
- Random walk can get stuck in dead ends.
- If there are dead ends, long-term visit rates are not well-defined (or non-sensical).

Teleporting – to get us of dead ends

# Teleporting – to get us of dead ends

- At a **dead end**, jump to a random web page with prob.  $1/N$ .

# Teleporting – to get us of dead ends

- At a **dead end**, jump to a random web page with prob.  $1/N$ .
- At a **non-dead end**, with probability 10%, jump to a random web page (to each with a probability of  $0.1/N$ ).

# Teleporting – to get us of dead ends

- At a **dead end**, jump to a random web page with prob.  $1/N$ .
- At a **non-dead end**, with probability 10%, jump to a random web page (to each with a probability of  $0.1/N$ ).
- With remaining probability (90%), go out on a random hyperlink.

# Teleporting – to get us of dead ends

- At a **dead end**, jump to a random web page with prob.  $1/N$ .
- At a **non-dead end**, with probability 10%, jump to a random web page (to each with a probability of  $0.1/N$ ).
- With remaining probability (90%), go out on a random hyperlink.
  - For example, if the page has 4 outgoing links: randomly choose one with probability  $(1-0.10)/4=0.225$

# Teleporting – to get us of dead ends

- At a **dead end**, jump to a random web page with prob.  $1/N$ .
- At a **non-dead end**, with probability 10%, jump to a random web page (to each with a probability of  $0.1/N$ ).
- With remaining probability (90%), go out on a random hyperlink.
  - For example, if the page has 4 outgoing links: randomly choose one with probability  $(1-0.10)/4=0.225$
- 10% is a parameter, the **teleportation rate**.



# Teleporting – to get us of dead ends

- At a **dead end**, jump to a random web page with prob.  $1/N$ .
- At a **non-dead end**, with probability 10%, jump to a random web page (to each with a probability of  $0.1/N$ ).
- With remaining probability (90%), go out on a random hyperlink.
  - For example, if the page has 4 outgoing links: randomly choose one with probability  $(1-0.10)/4=0.225$
- 10% is a parameter, the **teleportation rate**.
- Note: “jumping” from dead end is independent of teleportation rate.

# Result of teleporting

# Result of teleporting

- With teleporting, we cannot get stuck in a dead end.

# Result of teleporting

- With teleporting, we cannot get stuck in a dead end.
- But even without dead ends, a graph may not have well-defined long-term visit rates.

# Result of teleporting

- With teleporting, we cannot get stuck in a dead end.
- But even without dead ends in the original graph, we may not have well-defined long-term visit rates.
- More generally, we require that the Markov chain be **ergodic**.

# Ergodic Markov chains

- A Markov chain is ergodic if it is irreducible and aperiodic.

# Ergodic Markov chains

- A Markov chain is ergodic if it is irreducible and aperiodic.
- **Irreducibility**. Roughly: there is a path from any other page.

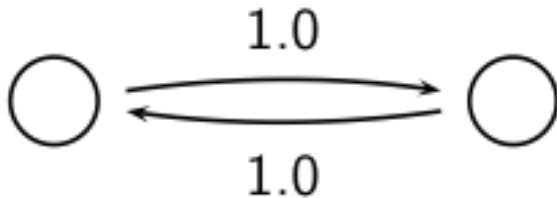
# Ergodic Markov chains

- A Markov chain is ergodic if it is irreducible and aperiodic.
- **Irreducibility**. Roughly: there is a path from any other page.
- **Aperiodicity**. Roughly: The pages cannot be partitioned such that the random walker visits the partitions sequentially.



# Ergodic Markov chains

- A Markov chain is ergodic if it is irreducible and aperiodic.
- **Irreducibility**. Roughly: there is a path from any other page.
- **Aperiodicity**. Roughly: The pages cannot be partitioned such that the random walker visits the partitions sequentially.
- A non-ergodic Markov chain:



# Ergodic Markov chains

- Theorem: For any ergodic Markov chain, there is a unique long-term visit rate for each state.

# Ergodic Markov chains

- Theorem: For any ergodic Markov chain, there is a unique long-term visit rate for each state.
- This is the **steady-state probability distribution**.

# Ergodic Markov chains

- Theorem: For any ergodic Markov chain, there is a unique long-term visit rate for each state.
- This is the **steady-state probability distribution**.
- Over a long time period, we visit each state in proportion to this rate.

# Ergodic Markov chains

- Theorem: For any ergodic Markov chain, there is a unique long-term visit rate for each state.
- This is the **steady-state probability distribution**.
- Over a long time period, we visit each state in proportion to this rate.
- It doesn't matter where we start.

# Ergodic Markov chains

- Theorem: For any ergodic Markov chain, there is a unique long-term visit rate for each state.
- This is the **steady-state probability distribution**.
- Over a long time period, we visit each state in proportion to this rate.
- It doesn't matter where we start.
- **Teleporting makes the web graph ergodic.**

# Ergodic Markov chains

- Theorem: For any ergodic Markov chain, there is a unique long-term visit rate for each state.
- This is the **steady-state probability distribution**.
- Over a long time period, we visit each state in proportion to this rate.
- It doesn't matter where we start.
- **Teleporting makes the web graph ergodic.**
- **$\implies$  Web-graph+teleporting has a steady-state probability distribution.**

# Ergodic Markov chains

- Theorem: For any ergodic Markov chain, there is a unique long-term visit rate for each state.
- This is the **steady-state probability distribution**.
- Over a long time period, we visit each state in proportion to this rate.
- It doesn't matter where we start.
- **Teleporting makes the web graph ergodic.**
- **$\implies$  Web-graph+teleporting has a steady-state probability distribution.**
- **$\implies$  Each page in the web-graph+teleporting has a PageRank.**



# Formalization of “visit”: Probability vector

# Formalization of “visit”: Probability vector

- A probability (row) vector  $\vec{x} = (x_1, \dots, x_N)$  tells us where the random walk is at any point.

# Formalization of “visit”: Probability vector

- A probability (row) vector  $\vec{x} = (x_1, \dots, x_N)$  tells us where the random walk is at any point.
- Example: 

(	0	0	0	...	1	...	0	0	0	)
	1	2	3	...	<i>i</i>	...	N-2	N-1	N	

# Formalization of “visit”: Probability vector

- A probability (row) vector  $\vec{x} = (x_1, \dots, x_N)$  tells us where the random walk is at any point.
- Example: 
$$\left( \begin{array}{cccccccccc} 0 & 0 & 0 & \dots & 1 & \dots & 0 & 0 & 0 \\ & 1 & 2 & 3 & \dots & i & \dots & N-2 & N-1 & N \end{array} \right)$$
- More generally: the random walk is on the page  $i$  with probability  $x_i$ .

# Formalization of “visit”: Probability vector

- A probability (row) vector  $\vec{x} = (x_1, \dots, x_N)$  tells us where the random walk is at any point.
- Example: 
$$\left( \begin{array}{cccccccccc} 0 & 0 & 0 & \dots & 1 & \dots & 0 & 0 & 0 \\ 1 & 2 & 3 & \dots & i & \dots & N-2 & N-1 & N \end{array} \right)$$
- More generally: the random walk is on the page  $i$  with probability  $x_i$ .
- Example: 
$$\left( \begin{array}{cccccccccc} 0.05 & 0.01 & 0.0 & \dots & 0.2 & \dots & 0.01 & 0.05 & 0.03 \\ 1 & 2 & 3 & \dots & i & \dots & N-2 & N-1 & N \end{array} \right)$$

# Formalization of “visit”: Probability vector

- A probability (row) vector  $\vec{x} = (x_1, \dots, x_N)$  tells us where the random walk is at any point.
- Example: 
$$\left( \begin{array}{cccccccccc} 0 & 0 & 0 & \dots & 1 & \dots & 0 & 0 & 0 \\ 1 & 2 & 3 & \dots & i & \dots & N-2 & N-1 & N \end{array} \right)$$
- More generally: the random walk is on the page  $i$  with probability  $x_i$ .
- Example: 
$$\left( \begin{array}{cccccccccc} 0.05 & 0.01 & 0.0 & \dots & 0.2 & \dots & 0.01 & 0.05 & 0.03 \\ 1 & 2 & 3 & \dots & i & \dots & N-2 & N-1 & N \end{array} \right)$$
- $\sum x_i = 1$

# Change in probability vector

- If the probability vector is  $\vec{x} = (x_1, \dots, x_N)$ , at this step, what is it at the next step?

# Change in probability vector

- If the probability vector is  $\vec{x} = (x_1, \dots, x_N)$ , at this step, what is it at the next step?
- Recall that row  $i$  of the transition probability matrix  $P$  tells us where we go next from state  $i$ .



# Change in probability vector

- If the probability vector is  $\vec{x} = (x_1, \dots, x_N)$ , at this step, what is it at the next step?
- Recall that row  $i$  of the transition probability matrix  $P$  tells us where we go next from state  $i$ .
- So from  $\vec{x}$ , our next state is distributed as  $\vec{x}P$ .

# Steady state in vector notation

# Steady state in vector notation

- The steady state in vector notation is simply a vector  $\vec{\pi} = (\pi_1, \pi_2, \dots, \pi_N)$  of probabilities.

# Steady state in vector notation

- The steady state in vector notation is simply a vector  $\vec{\pi} = (\pi_1, \pi_2, \dots, \pi_N)$  of probabilities.
- (We use  $\vec{\pi}$  to distinguish it from the notation for the probability vector  $\vec{x}$ .)

# Steady state in vector notation

- The steady state in vector notation is simply a vector  $\vec{\pi} = (\pi_1, \pi_2, \dots, \pi_N)$  of probabilities.
- (We use  $\vec{\pi}$  to distinguish it from the notation for the probability vector  $\vec{x}$ .)
- $\pi$  is the long-term visit rate (or PageRank) of page  $i$ .

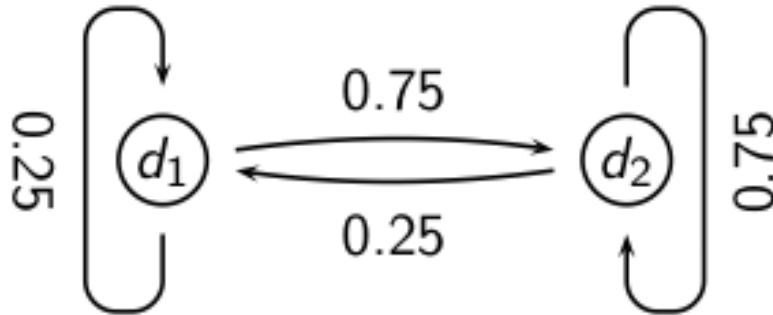
# Steady state in vector notation

- The steady state in vector notation is simply a vector  $\vec{\pi} = (\pi_1, \pi_2, \dots, \pi_N)$  of probabilities.
- (We use  $\vec{\pi}$  to distinguish it from the notation for the probability vector  $\vec{x}$ .)
- $\pi$  is the long-term visit rate (or PageRank) of page  $i$ .
- So we can think of PageRank as a very long vector – one entry per page.

# Steady-state distribution: Example

# Steady-state distribution: Example

- What is the PageRank / steady state in this example?





# Steady-state distribution: Example

# Steady-state distribution: Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.25$	$P_{12} = 0.75$
			$P_{21} = 0.25$	$P_{22} = 0.75$
$t_0$	0.25	0.75		
$t_1$				

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Steady-state distribution: Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.25$	$P_{12} = 0.75$
			$P_{21} = 0.25$	$P_{22} = 0.75$
$t_0$	0.25	0.75	0.25	0.75
$t_1$				

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Steady-state distribution: Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.25$	$P_{12} = 0.75$
			$P_{21} = 0.25$	$P_{22} = 0.75$
$t_0$	0.25	0.75	0.25	0.75
$t_1$	0.25	0.75		

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Steady-state distribution: Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.25$	$P_{12} = 0.75$
			$P_{21} = 0.25$	$P_{22} = 0.75$
$t_0$	0.25	0.75	0.25	0.75
$t_1$	0.25	0.75	(convergence)	

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

How do we compute the steady state vector?

# How do we compute the steady state vector?

- In other words: how do we compute PageRank?

# How do we compute the steady state vector?

- In other words: how do we compute PageRank?
- Recall:  $\vec{\pi} = (\pi_1, \pi_2, \dots, \pi_N)$  is the PageRank vector, the vector of steady-state probabilities ...



# How do we compute the steady state vector?

- In other words: how do we compute PageRank?
- Recall:  $\vec{\pi} = (\pi_1, \pi_2, \dots, \pi_N)$  is the PageRank vector, the vector of steady-state probabilities ...
- ... and if the distribution in this step is  $\vec{x}$ , then the distribution in the next step is  $\vec{x}P$ .

# How do we compute the steady state vector?

- In other words: how do we compute PageRank?
- Recall:  $\vec{\pi} = (\pi_1, \pi_2, \dots, \pi_N)$  is the PageRank vector, the vector of steady-state probabilities ...
- ... and if the distribution in this step is  $\vec{x}$ , then the distribution in the next step is  $\vec{x}P$ .
- But  $\vec{\pi}$  is the steady state!

# How do we compute the steady state vector?

- In other words: how do we compute PageRank?
- Recall:  $\vec{\pi} = (\pi_1, \pi_2, \dots, \pi_N)$  is the PageRank vector, the vector of steady-state probabilities ...
- ... and if the distribution in this step is  $\vec{x}$ , then the distribution in the next step is  $\vec{x}P$ .
- But  $\vec{\pi}$  is the steady state!
- So:  $\vec{\pi} = \vec{\pi} P$

# How do we compute the steady state vector?

- In other words: how do we compute PageRank?
- Recall:  $\vec{\pi} = (\pi_1, \pi_2, \dots, \pi_N)$  is the PageRank vector, the vector of steady-state probabilities ...
- ... and if the distribution in this step is  $\vec{x}$ , then the distribution in the next step is  $\vec{x}P$ .
- But  $\vec{\pi}$  is the steady state!
- So:  $\vec{\pi} = \vec{\pi} P$
- Solving this matrix equation gives us  $\vec{\pi}$ .

# How do we compute the steady state vector?

- In other words: how do we compute PageRank?
- Recall:  $\vec{\pi} = (\pi_1, \pi_2, \dots, \pi_N)$  is the PageRank vector, the vector of steady-state probabilities ...
- ... and if the distribution in this step is  $\vec{x}$ , then the distribution in the next step is  $\vec{x}P$ .
- But  $\vec{\pi}$  is the steady state!
- So:  $\vec{\pi} = \vec{\pi} P$
- Solving this matrix equation gives us  $\vec{\pi}$ .
- $\vec{\pi}$  is the principal left eigenvector for  $P$  ...

# How do we compute the steady state vector?

- In other words: how do we compute PageRank?
- Recall:  $\vec{\pi} = (\pi_1, \pi_2, \dots, \pi_N)$  is the PageRank vector, the vector of steady-state probabilities ...
- ... and if the distribution in this step is  $\vec{x}$ , then the distribution in the next step is  $\vec{x}P$ .
- But  $\vec{\pi}$  is the steady state!
- So:  $\vec{\pi} = \vec{\pi} P$
- Solving this matrix equation gives us  $\vec{\pi}$ .
- $\vec{\pi}$  is the principal left eigenvector for  $P$  ...
- ... that is,  $\vec{\pi}$  is the left eigenvector with the largest eigenvalue.

# How do we compute the steady state vector?

- In other words: how do we compute PageRank?
- Recall:  $\vec{\pi} = (\pi_1, \pi_2, \dots, \pi_N)$  is the PageRank vector, the vector of steady-state probabilities ...
- ... and if the distribution in this step is  $\vec{x}$ , then the distribution in the next step is  $\vec{x}P$ .
- But  $\vec{\pi}$  is the steady state!
- So:  $\vec{\pi} = \vec{\pi} P$
- Solving this matrix equation gives us  $\vec{\pi}$ .
- $\vec{\pi}$  is the principal left eigenvector for  $P$  ...
- ... that is,  $\vec{\pi}$  is the left eigenvector with the largest eigenvalue.
- All transition probability matrices have largest eigenvalue 1.

One way of computing the PageRank  $\vec{\pi}$



# One way of computing the PageRank $\vec{\pi}$

- Start with any distribution  $\vec{x}$ , e.g., uniform distribution

# One way of computing the PageRank $\vec{\pi}$

- Start with any distribution  $\vec{x}$ , e.g., uniform distribution
- After one step, we're at  $\vec{x}P$ .

# One way of computing the PageRank $\vec{\pi}$

- Start with any distribution  $\vec{x}$ , e.g., uniform distribution
- After one step, we're at  $\vec{x}P$ .
- After two steps, we're at  $\vec{x}P^2$ .

# One way of computing the PageRank $\vec{\pi}$

- Start with any distribution  $\vec{x}$ , e.g., uniform distribution
- After one step, we're at  $\vec{x}P$ .
- After two steps, we're at  $\vec{x}P^2$ .
- After  $k$  steps, we're at  $\vec{x}P^k$ .

# One way of computing the PageRank $\vec{\pi}$

- Start with any distribution  $\vec{x}$ , e.g., uniform distribution
- After one step, we're at  $\vec{x}P$ .
- After two steps, we're at  $\vec{x}P^2$ .
- After  $k$  steps, we're at  $\vec{x}P^k$ .
- Algorithm: multiply  $\vec{x}$  by increasing powers of  $P$  until convergence.

# One way of computing the PageRank $\vec{\pi}$

- Start with any distribution  $\vec{x}$ , e.g., uniform distribution
- After one step, we're at  $\vec{x}P$ .
- After two steps, we're at  $\vec{x}P^2$ .
- After  $k$  steps, we're at  $\vec{x}P^k$ .
- Algorithm: multiply  $\vec{x}$  by increasing powers of  $P$  until convergence.
- This is called the **power method**.

# One way of computing the PageRank $\vec{\pi}$

- Start with any distribution  $\vec{x}$ , e.g., uniform distribution
- After one step, we're at  $\vec{x}P$ .
- After two steps, we're at  $\vec{x}P^2$ .
- After  $k$  steps, we're at  $\vec{x}P^k$ .
- Algorithm: multiply  $\vec{x}$  by increasing powers of  $P$  until convergence.
- This is called the **power method**.
- Recall: regardless of where we start, we eventually reach the steady state  $\vec{\pi}$ .

# One way of computing the PageRank $\vec{\pi}$

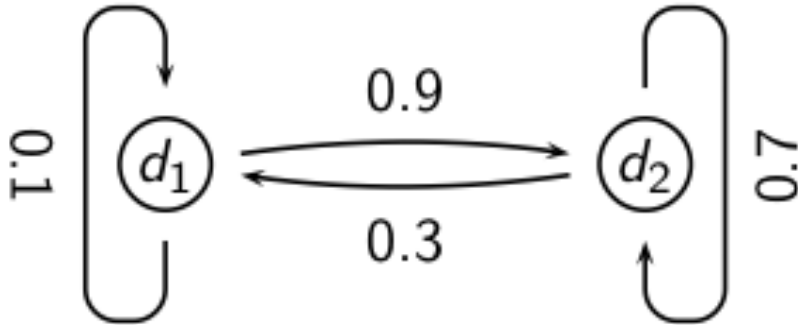
- Start with any distribution  $\vec{x}$ , e.g., uniform distribution
- After one step, we're at  $\vec{x}P$ .
- After two steps, we're at  $\vec{x}P^2$ .
- After  $k$  steps, we're at  $\vec{x}P^k$ .
- Algorithm: multiply  $\vec{x}$  by increasing powers of  $P$  until convergence.
- This is called the **power method**.
- Recall: regardless of where we start, we eventually reach the steady state  $\vec{\pi}$ .
- Thus: we will eventually (in asymptotia) reach the steady state.



# Power method: Example

# Power method: Example

- What is the PageRank / steady state in this example?



# Computing PageRank: Power Example

# Computing PageRank: Power Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.1$	$P_{12} = 0.9$
			$P_{21} = 0.3$	$P_{22} = 0.7$
$t_0$	0	1		$= \vec{x}P$
$t_1$				$= \vec{x}P^2$
$t_2$				$= \vec{x}P^3$
$t_3$				$= \vec{x}P^4$
				$\dots$
$t_\infty$				$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Computing PageRank: Power Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$			
			$P_{11} = 0.1$	$P_{12} = 0.9$	
			$P_{21} = 0.3$	$P_{22} = 0.7$	
$t_0$	0	1	0.3	0.7	$= \vec{x}P$
$t_1$					$= \vec{x}P^2$
$t_2$					$= \vec{x}P^3$
$t_3$					$= \vec{x}P^4$
					$\dots$
$t_\infty$					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Computing PageRank: Power Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$			
			$P_{11} = 0.1$	$P_{12} = 0.9$	
			$P_{21} = 0.3$	$P_{22} = 0.7$	
$t_0$	0	1	0.3	0.7	$= \vec{x}P$
$t_1$	0.3	0.7			$= \vec{x}P^2$
$t_2$					$= \vec{x}P^3$
$t_3$					$= \vec{x}P^4$
					$\dots$
$t_\infty$					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Computing PageRank: Power Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$			
			$P_{11} = 0.1$	$P_{12} = 0.9$	
			$P_{21} = 0.3$	$P_{22} = 0.7$	
$t_0$	0	1	0.3	0.7	$= \vec{x}P$
$t_1$	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
$t_2$					$= \vec{x}P^3$
$t_3$					$= \vec{x}P^4$
					$\dots$
$t_\infty$					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Computing PageRank: Power Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$			
			$P_{11} = 0.1$	$P_{12} = 0.9$	
			$P_{21} = 0.3$	$P_{22} = 0.7$	
$t_0$	0	1	0.3	0.7	$= \vec{x}P$
$t_1$	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
$t_2$	0.24	0.76			$= \vec{x}P^3$
$t_3$					$= \vec{x}P^4$
					$\dots$
$t_\infty$					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$



# Computing PageRank: Power Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$			
			$P_{11} = 0.1$	$P_{12} = 0.9$	
			$P_{21} = 0.3$	$P_{22} = 0.7$	
$t_0$	0	1	0.3	0.7	$= \vec{x}P$
$t_1$	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
$t_2$	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
$t_3$					$= \vec{x}P^4$
					$\dots$
$t_\infty$					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Computing PageRank: Power Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$			
			$P_{11} = 0.1$	$P_{12} = 0.9$	
			$P_{21} = 0.3$	$P_{22} = 0.7$	
$t_0$	0	1	0.3	0.7	$= \vec{x}P$
$t_1$	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
$t_2$	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
$t_3$	0.252	0.748			$= \vec{x}P^4$
					$\dots$
$t_\infty$					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Computing PageRank: Power Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$			
			$P_{11} = 0.1$	$P_{12} = 0.9$	
			$P_{21} = 0.3$	$P_{22} = 0.7$	
$t_0$	0	1	0.3	0.7	$= \vec{x}P$
$t_1$	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
$t_2$	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
$t_3$	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
					$\dots$
$t_\infty$					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Computing PageRank: Power Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$			
			$P_{11} = 0.1$	$P_{12} = 0.9$	
			$P_{21} = 0.3$	$P_{22} = 0.7$	
$t_0$	0	1	0.3	0.7	$= \vec{x}P$
$t_1$	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
$t_2$	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
$t_3$	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
				...	...
$t_\infty$					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Computing PageRank: Power Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$			
			$P_{11} = 0.1$	$P_{12} = 0.9$	
			$P_{21} = 0.3$	$P_{22} = 0.7$	
$t_0$	0	1	0.3	0.7	$= \vec{x}P$
$t_1$	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
$t_2$	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
$t_3$	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
				...	...
$t_\infty$	0.25	0.75			$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Computing PageRank: Power Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$			
			$P_{11} = 0.1$	$P_{12} = 0.9$	
			$P_{21} = 0.3$	$P_{22} = 0.7$	
$t_0$	0	1	0.3	0.7	$= \vec{x}P$
$t_1$	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
$t_2$	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
$t_3$	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
				...	...
$t_\infty$	0.25	0.75	0.25	0.75	$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Computing PageRank: Power Example

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$			
			$P_{11} = 0.1$	$P_{12} = 0.9$	
			$P_{21} = 0.3$	$P_{22} = 0.7$	
$t_0$	0	1	0.3	0.7	$= \vec{x}P$
$t_1$	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
$t_2$	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
$t_3$	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
				...	...
$t_\infty$	0.25	0.75	0.25	0.75	$= \vec{x}P^\infty$

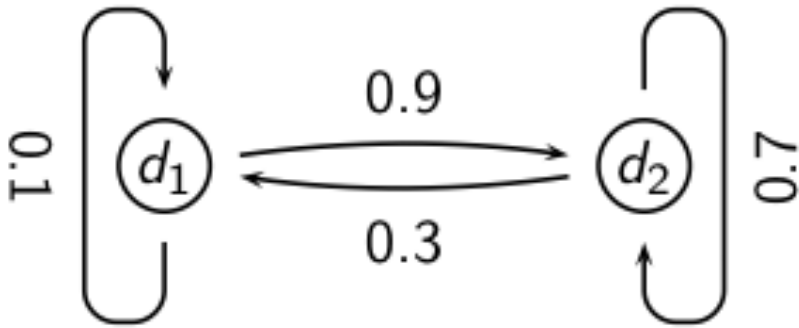
PageRank vector  $= \vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Power method: Example

- What is the PageRank / steady state in this example?

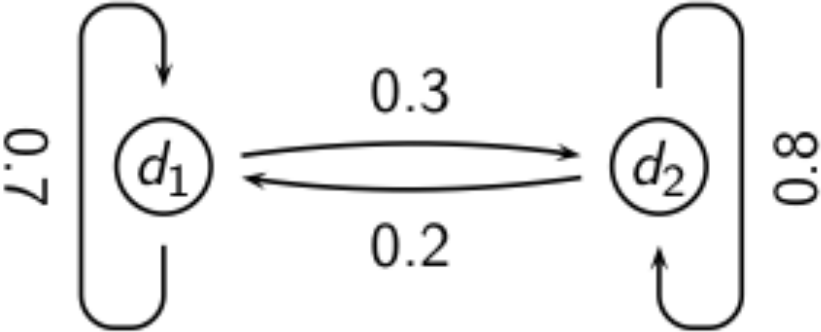


- The steady state distribution (= the PageRanks) in this example are 0.25 for  $d_1$  and 0.75 for  $d_2$ .



Exercise: Compute PageRank using power method

# Exercise: Compute PageRank using power method



# Solution

# Solution

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
$t_0$	0	1		
$t_1$				
$t_2$				
$t_3$				
$t_\infty$				

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Solution

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
$t_0$	0	1	0.2	0.8
$t_1$				
$t_2$				
$t_3$				
$t_\infty$				

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Solution

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
$t_0$	0	1	0.2	0.8
$t_1$	0.2	0.8		
$t_2$				
$t_3$				
$t_\infty$				

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Solution

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
$t_0$	0	1	0.2	0.8
$t_1$	0.2	0.8	0.3	0.7
$t_2$				
$t_3$				
$t_\infty$				

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Solution

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
$t_0$	0	1	0.2	0.8
$t_1$	0.2	0.8	0.3	0.7
$t_2$	0.3	0.7		
$t_3$				
$t_\infty$				

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$



# Solution

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
$t_0$	0	1	0.2	0.8
$t_1$	0.2	0.8	0.3	0.7
$t_2$	0.3	0.7	0.35	0.65
$t_3$				
$t_\infty$				

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Solution

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
$t_0$	0	1	0.2	0.8
$t_1$	0.2	0.8	0.3	0.7
$t_2$	0.3	0.7	0.35	0.65
$t_3$	0.35	0.65		
$t_\infty$				

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Solution

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
$t_0$	0	1	0.2	0.8
$t_1$	0.2	0.8	0.3	0.7
$t_2$	0.3	0.7	0.35	0.65
$t_3$	0.35	0.65	0.375	0.625
$t_\infty$				

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Solution

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
$t_0$	0	1	0.2	0.8
$t_1$	0.2	0.8	0.3	0.7
$t_2$	0.3	0.7	0.35	0.65
$t_3$	0.35	0.65	0.375	0.625
			...	
$t_\infty$				

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Solution

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
$t_0$	0	1	0.2	0.8
$t_1$	0.2	0.8	0.3	0.7
$t_2$	0.3	0.7	0.35	0.65
$t_3$	0.35	0.65	0.375	0.625
			...	
$t_\infty$	0.4	0.6		

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# Solution

	$x_1$ $P_t(d_1)$	$x_2$ $P_t(d_2)$		
			$P_{11} = 0.7$	$P_{12} = 0.3$
			$P_{21} = 0.2$	$P_{22} = 0.8$
$t_0$	0	1	0.2	0.8
$t_1$	0.2	0.8	0.3	0.7
$t_2$	0.3	0.7	0.35	0.65
$t_3$	0.35	0.65	0.375	0.625
			...	
$t_\infty$	0.4	0.6	0.4	0.6

PageRank vector =  $\vec{\pi} = (\pi_1, \pi_2) = (0.4, 0.6)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

# PageRank summary

# PageRank summary

- Preprocessing



# PageRank summary

- Preprocessing
  - Given graph of links, build matrix  $P$

# PageRank summary

- Preprocessing
  - Given graph of links, build matrix  $P$
  - Apply teleportation

# PageRank summary

- Preprocessing
  - Given graph of links, build matrix  $P$
  - Apply teleportation
  - From modified matrix, compute  $\vec{\pi}$

# PageRank summary

- Preprocessing
  - Given graph of links, build matrix  $P$
  - Apply teleportation
  - From modified matrix, compute  $\vec{\pi}$
  - $\pi_i$  is the PageRank of page  $i$ .

# PageRank summary

- Preprocessing
  - Given graph of links, build matrix  $P$
  - Apply teleportation
  - From modified matrix, compute  $\vec{\pi}$
  - $\vec{\pi}_i$  is the PageRank of page  $i$ .
- Query processing

# PageRank summary

- Preprocessing
  - Given graph of links, build matrix  $P$
  - Apply teleportation
  - From modified matrix, compute  $\vec{\pi}$
  - $\pi_i$  is the PageRank of page  $i$ .
- Query processing
  - Retrieve pages satisfying the query

# PageRank summary

- Preprocessing
  - Given graph of links, build matrix  $P$
  - Apply teleportation
  - From modified matrix, compute  $\vec{\pi}$
  - $\pi_i$  is the PageRank of page  $i$ .
- Query processing
  - Retrieve pages satisfying the query
  - Rank them by their PageRank

# PageRank summary

- Preprocessing
  - Given graph of links, build matrix  $P$
  - Apply teleportation
  - From modified matrix, compute  $\vec{\pi}$
  - $\vec{\pi}_i$  is the PageRank of page  $i$ .
- Query processing
  - Retrieve pages satisfying the query
  - Rank them by their PageRank
  - Return reranked list to the user



# PageRank issues

- Real surfers are not random surfers.
  - Examples of nonrandom surfing: back button, short vs. long paths, bookmarks, directories – and search!
  - → Markov model is not a good model of surfing.
  - But it's good enough as a model for our purposes.
- Simple PageRank ranking (as described on previous slide) produces bad results for many pages.
  - Consider the query [video service].
  - The Yahoo home page (i) has a very high PageRank and (ii) contains both *video* and *service*.
  - If we rank all pages containing the query terms according to PageRank, then the Yahoo home page would be top-ranked.
  - Clearly not desirable.

# How important is PageRank?

- Frequent claim: PageRank is the most important component of web ranking.
- The reality:
  - There are several components that are at least as important: e.g., anchor text, phrases, proximity, tiered indexes ...
  - Rumor has it that PageRank in his original form (as presented here) now has a negligible impact on ranking!
  - However, variants of a page's PageRank are still an essential part of ranking.
  - Addressing link spam is difficult and crucial.