PageRank

Debapriyo Majumdar

Indian Statistical Institute

debapriyo@isical.ac.in

Early search engines

- Main approach: matching terms (loosely speaking, words)
  - If query terms are present in a document, then the document is possibly relevant
  - Tf.IDF (and many other variants): assign a score for every term in a document
    - Intuition 1: more times a term is present in a document, the more important it is
      - Term frequency (TF)
    - Intuition 2: If a term is present in many documents, it is not *special* in any of the documents
      - (Inverse) document frequency (IDF)
  - Rank documents based on these scores
    - Higher Tf.IDF score $\Longrightarrow$ document showed up higher in search engine ranking

The spammer wants to exploit the ranking algorithm

- Spammers' goal: get their pages to show up in the search results to receive clicks
  - End goal: advertising, phishing, malware spreading, …
- How to get a page up in the search ranking?

Query: **amir khan movie**

pk amir
khan movie

movie
amir khan

amir khan amir
khan amir khan
buy this pay here
shahrukh khan

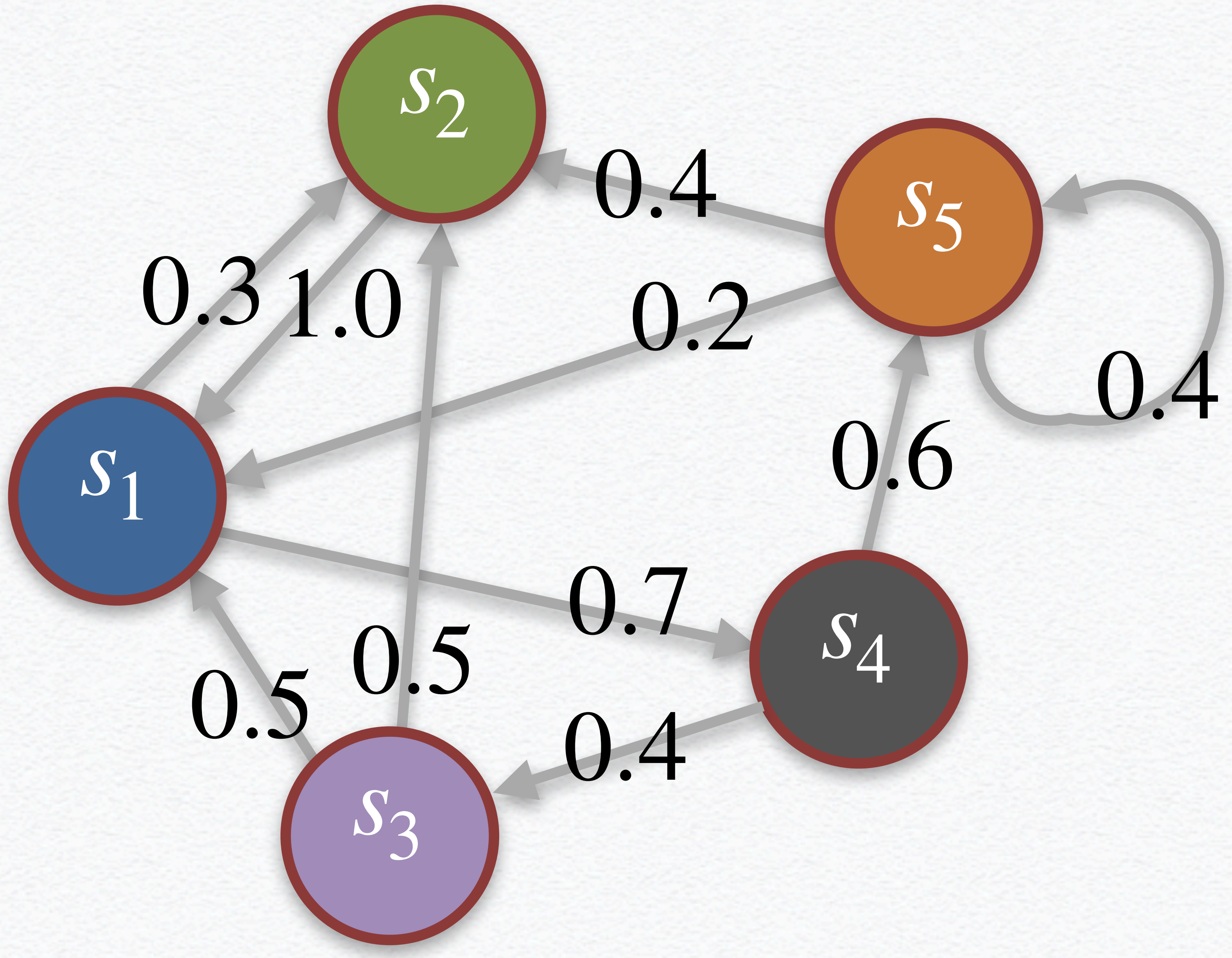
Term spam

I'll create documents with the popular query terms being present many times

PageRank

- Assumption (a reasonable one)
  - Users of the web are largely reasonable people
  - They put (more) links to useful pages
- PageRank
  - Named after Larry Page (co-founder of Google Inc.)
  - Patented by Stanford University, later bought by Google
- Approach
  - Importance (PageRank) of a webpage is influenced by the number and quality of links into the page
  - Search results ranked by term matching as well as PageRank
  - Intuition – Random web surfer model: a random surfer follows links and surfs the web. More likely to end up at more important pages
- Advantage: term spam cannot ensure in-links into those pages
- Many variations of PageRank

$n = 5$ in this example



	1.0	0.5		0.2
0.3		0.5		0.4
			0.4	
0.7				
			0.6	0.4

Markov Chain

- A **discrete-time stochastic** (random) process
- Set $\mathcal{S} = \{s_1, \ldots, s_N\}$ of n states
- In any one of the states at any given time step t
- A probability distribution $p : S \times S \rightarrow [0,1]$ determines the probabilities of going to a state at the next time step $t + 1$
- Can define a transition matrix $M = (p_{ij})_{1 \leq i,j \leq n}$ as $p_{ij} := p(s_i \,|\, s_j) = p[S_{t+1} = s_i \,|\, S_t = s_j]$
  - If at state s_j now, the probability of going to state s_i in the next step is p_{ij}
- Markov property: $\sum_{i=1}^{n} p_{ij} = 1$, for all $j = 1, \ldots, n$.

Markov Chain and Stochastic Matrix

- A matrix $M = (p_{ij})_{1 \le i,j \le n}$ with the following properties

  All entries represent probabilities: $p_{ij} \in [0,1]$, and

  Column sums are 1: $\sum_{i=1}^{n} p_{ij} = 1$, for all $j = 1,\ldots,n$.

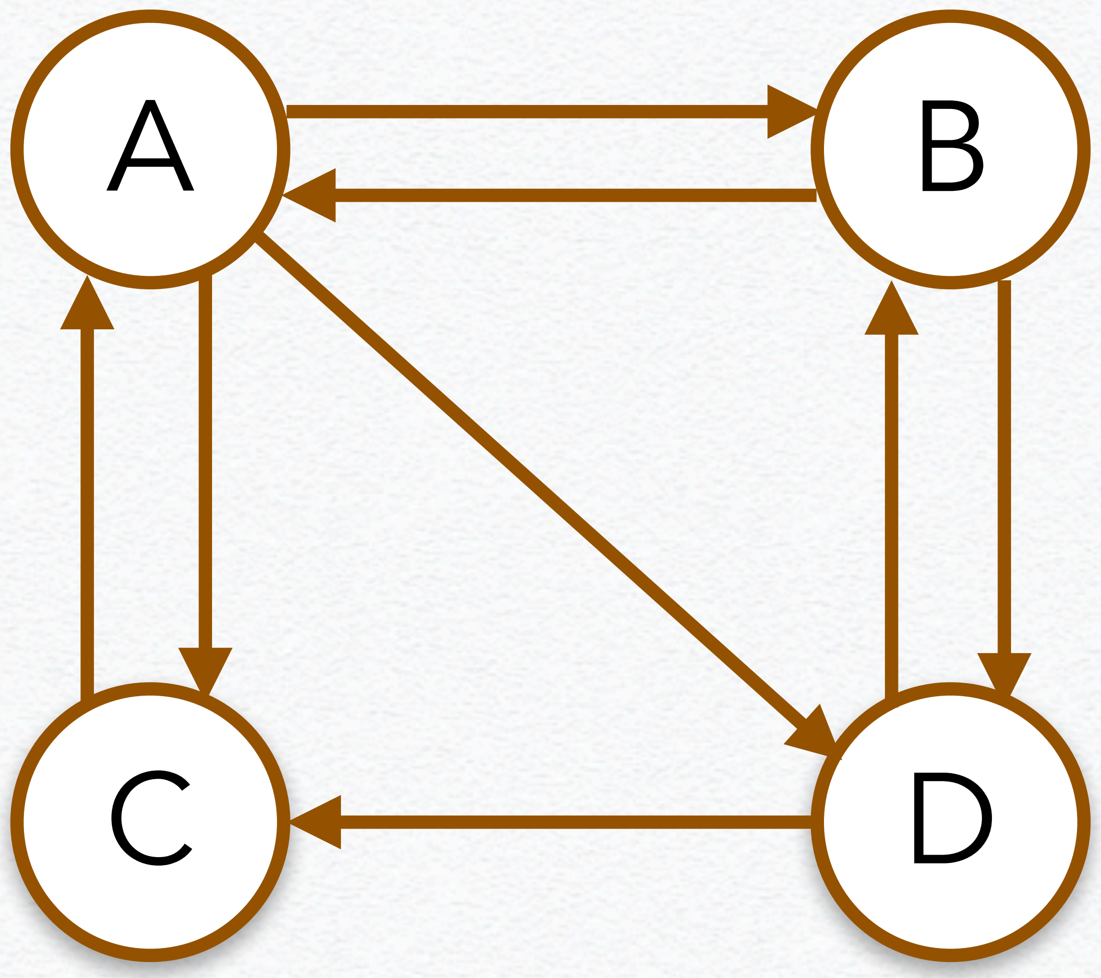
  Is called a **left stochastic matrix**.

- Note: the formulation is also valid with a right stochastic matrix (transpose), where the row sums are 1.
- Property of a left stochastic matrix:
  - Largest eigenvalue is 1.
  - $Mv = v$ where v is the principal eigenvector.

	1.0	0.5		0.2
0.3		0.5		0.4
			0.4	
0.7				
			0.6	0.4

The random surfer model

A tiny web



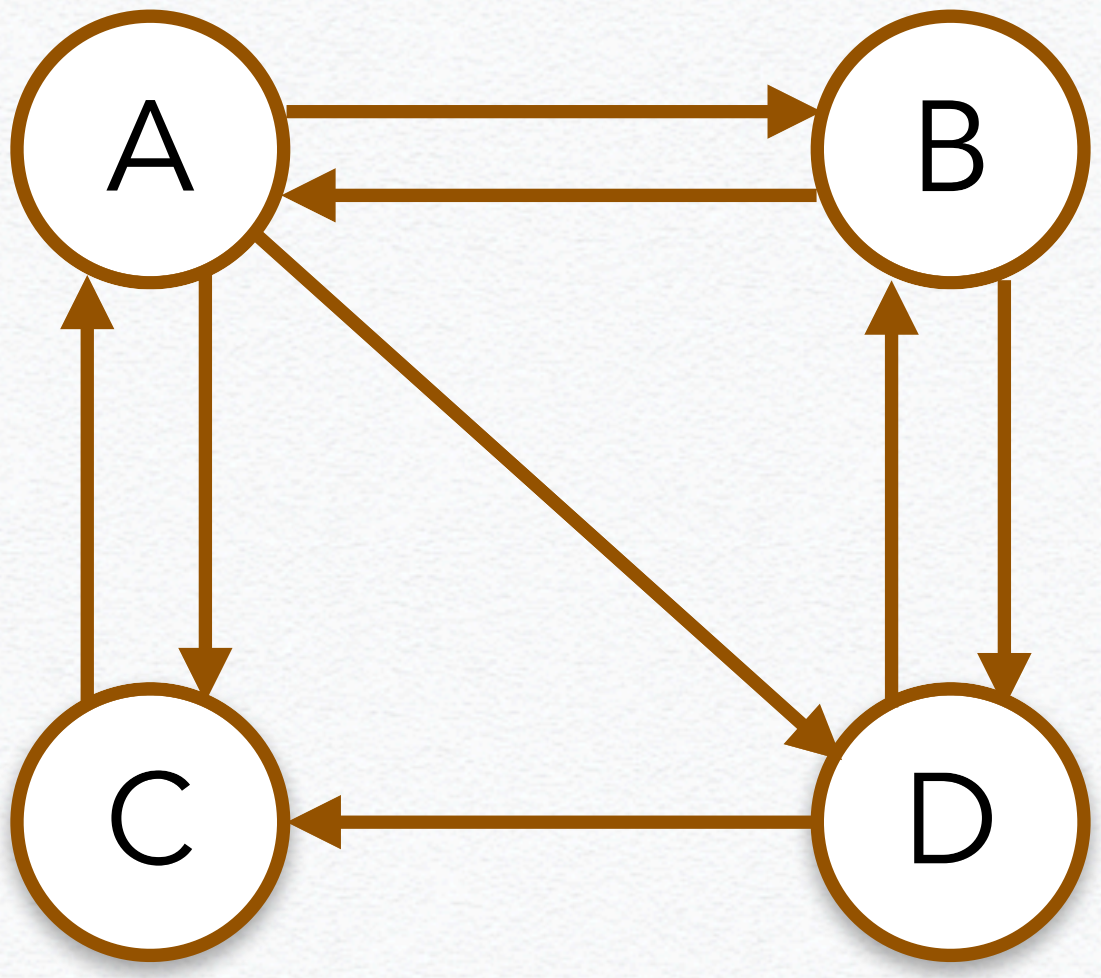
	A	B	C	D
	0	1/2	1	0
$M =$	1/3	0	0	1/2
	1/3	0	0	1/2
	1/3	1/2	0	0

- Web graph, links are directed edges
  - Assume equal weights in this example
  - If a surfer starts at A, with probability 1/3 each, may go to B, C, or D
  - If a surfer starts at B, with probability 1/2 each may go to A or D
  - Can define a transition matrix
- Markov process:
  - Future state solely based on present
  - $M_{ij} = P[j \rightarrow i$ in the next step | presently in $j]$

Example courtesy: book by Leskovec, Rajaraman and Ullman

The random surfer model

A tiny web



- Random surfer: initially at any position, with equal probability $1/n$
- Distribution (column) vector $v = (1/n, \ldots, 1/n)$
- Probability distribution for her location after one step?
- Distribution vector: Mv
- Distribution vector after 2 steps: $M^2 v$

Initially at A (1/4): A → A: not possible. Probability = 0
Initially at B (1/4): B → A (1/2). Overall probability = 1/8
Initially at C (1/4): C → A (1). Overall probability = 1/4
Initially at D (1/4): D →A (0). Overall probability = 0

$M =$

A	B	C	D
0	1/2	1	0
1/3	0	0	1/2
1/3	0	0	1/2
1/3	1/2	0	0

$v =$

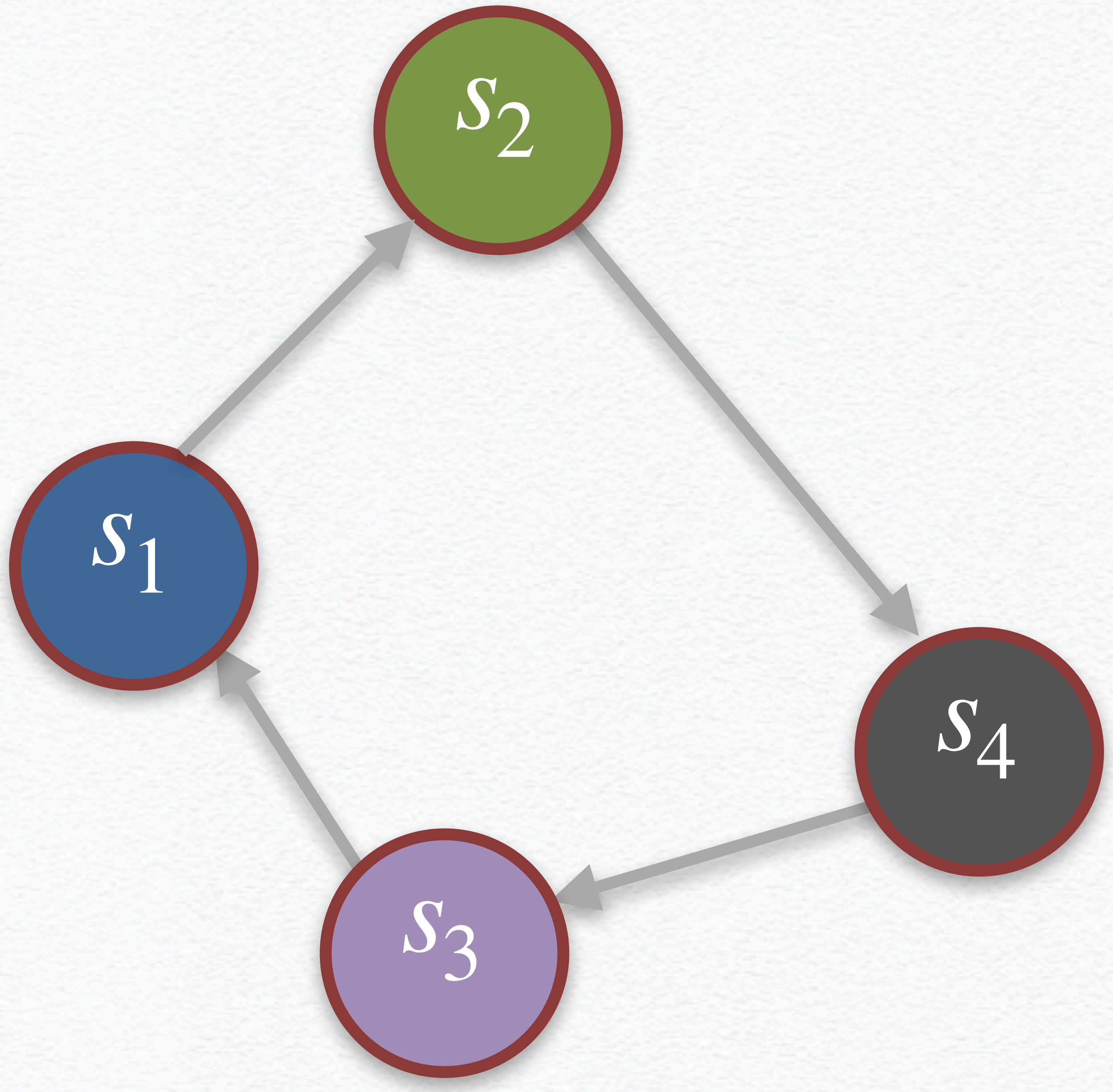
1/4
1/4
1/4
1/4

$Mv =$

0 + 1/8 + 1/4 + 0 = 9/24
1/12 + 0 + 0 + 1/8 = 5/24
1/12 + 0 + 0 + 1/8 = 5/24
1/12 + 1/8 + 0 + 0 = 5/24

Ergodic Markov Chain

- **Ergodic** Markov chain: if there exists a positive integer t_0 such that for all pairs of states s_i, s_j in the Markov chain, if it is started at time 0 in state s_i then for all $t > t_0$, the probability of being in state s_j at time t is greater than 0.
  - In other words, eventually, the probability of being at any state at any point of time is positive.
- Necessary conditions for ergodicity
  - (a) **Irreducibility**: there is a sequence of non-zero probability from any state to another.
  - (b) **Aperiodicity**: states are not partitioned into sets such that all transitions happen cyclically from one to another (not periodic).



Example of periodic Markov chain

Perron — Frobenius theorem (1912)

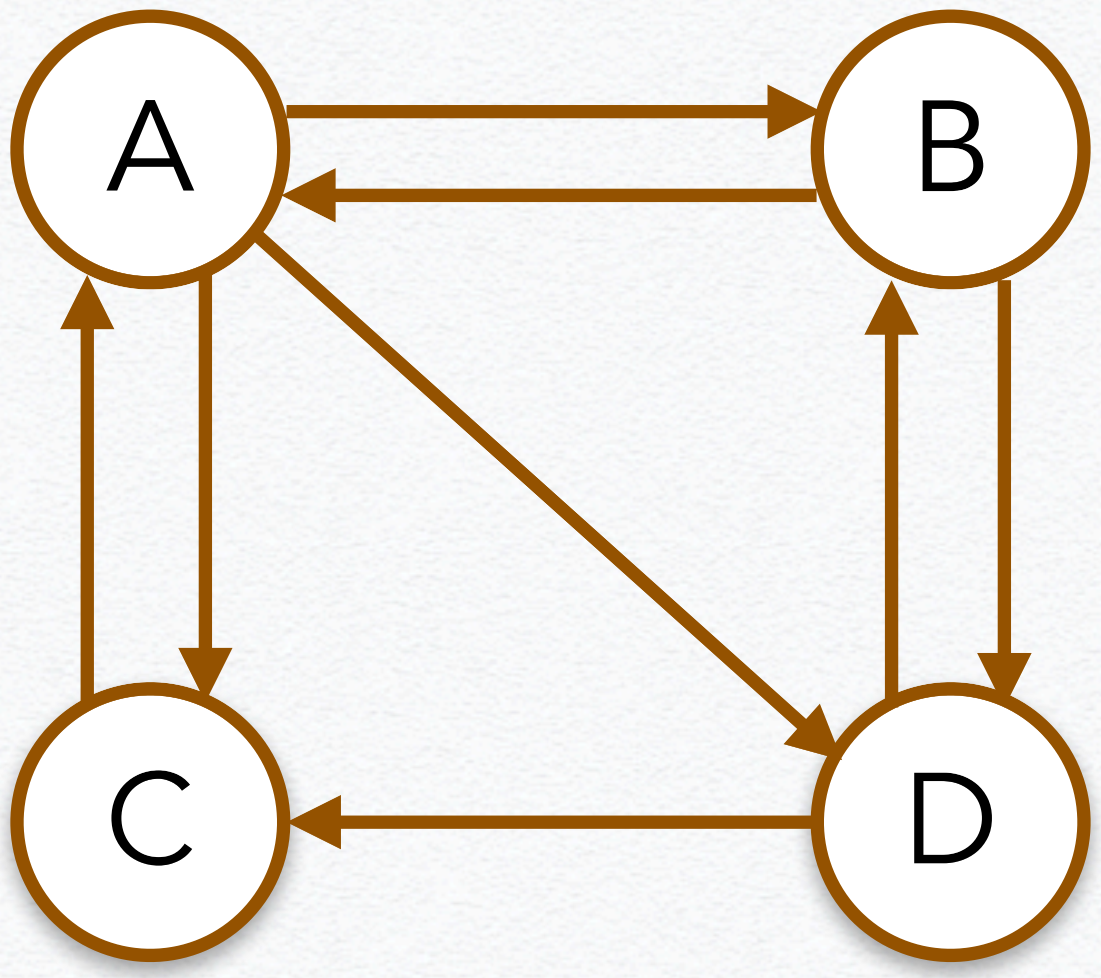
- For any ergodic Markov chain, For any ergodic Markov chain, there is a unique steady-state probability vector $\pi = [\pi_1, \pi_2, \ldots, \pi_n]^T$ that is the principal eigenvector of the transition matrix M, such that if $\eta(i, t)$is the number of visits to state s_i in t steps, then

$$\lim_{t \to \infty} \frac{\eta(i, t)}{t} = \pi_i$$

- In other words, the probability distribution converges to a limiting distribution π with $M\pi = \pi$

PageRank

A tiny web



- PageRank: the stationary distribution (column) vector π
- π_i is the probably of the random surfer being at state s_i eventually
- Computation:

  Initialize $v := (1/n, \ldots, 1/n)$

  while (norm($Mv - v$) > ε) {

  $\quad v := Mv$

  }

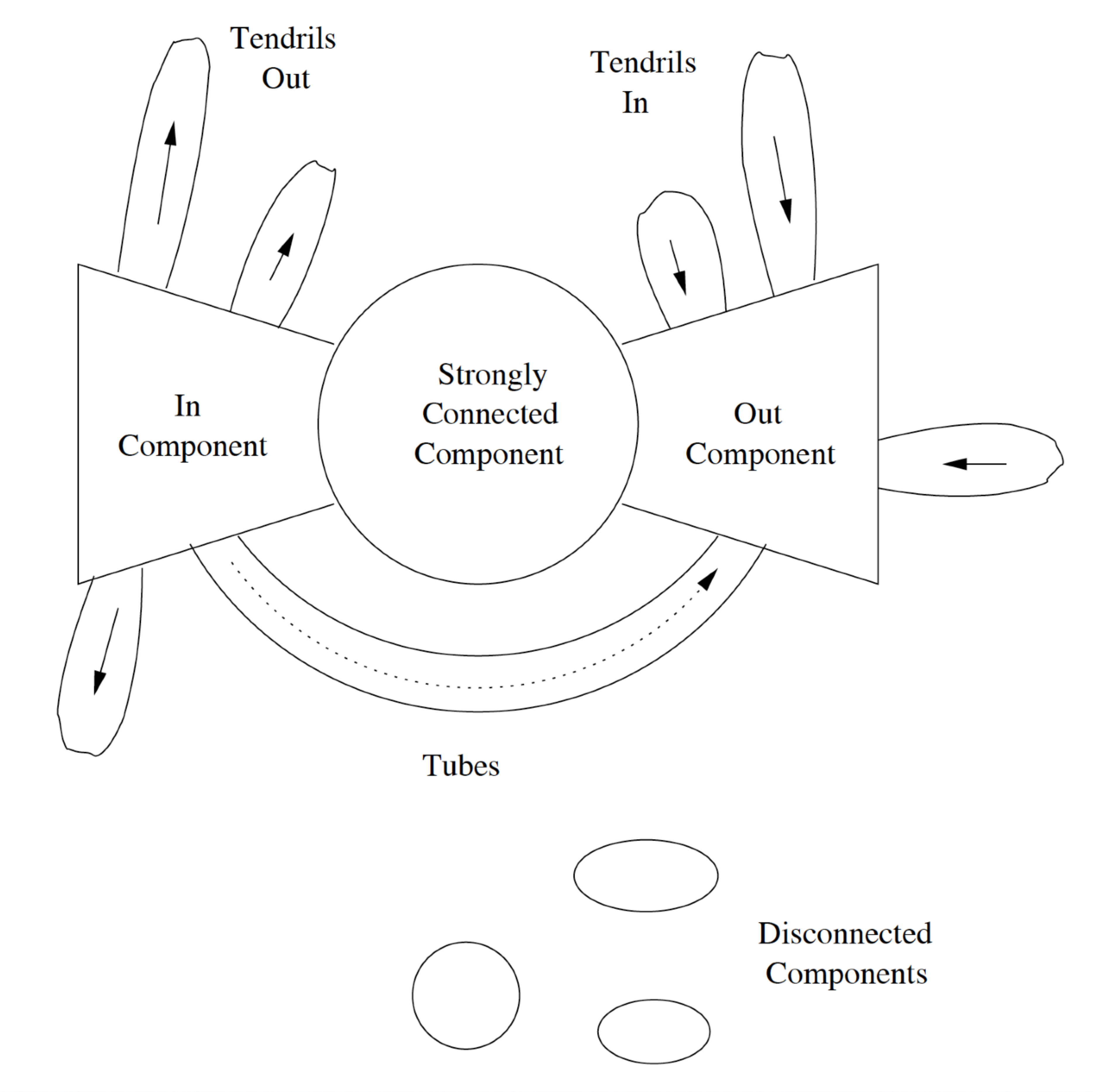
M

0	1/2	1	0
1/3	0	0	1/2
1/3	0	0	1/2
1/3	1/2	0	0

v	Mv	M^2v		M^kv	
1/4	9/24	15/48		3/9	PageRank of A
1/4	5/24	11/48		2/9	PageRank of B
1/4	5/24	11/48	$\cdots \longrightarrow$	2/9	PageRank of C
1/4	5/24	11/48		2/9	PageRank of D

Reality check: structure of the web

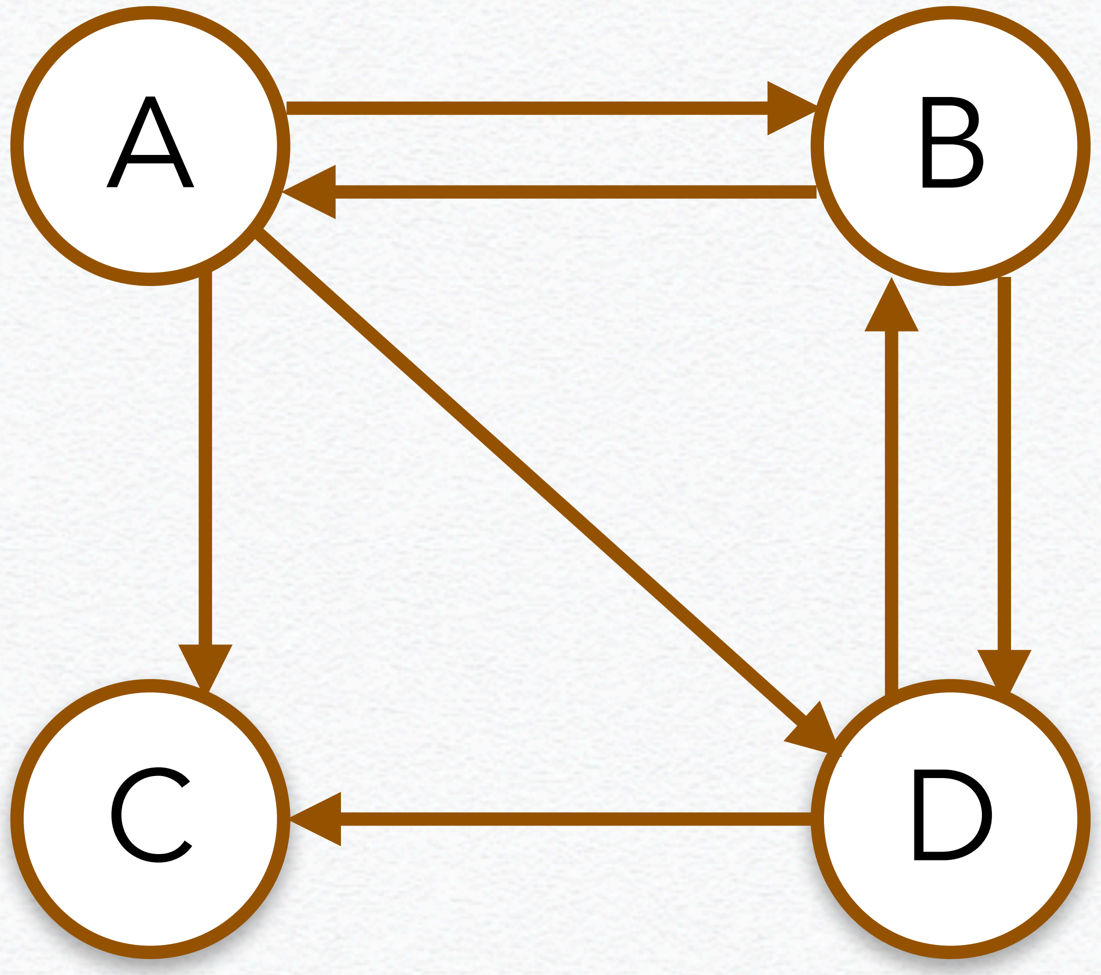
- The web is ***not*** strongly connected ☹
- The formulated Markov chain is not ergodic
- An early study of the web showed
  - One large strongly connected component
  - Several other components
- Requires modification to PageRank approach
- Two main problems
  1. Dead ends: a page with no outlink
  2. Spider traps: group of pages, outlinks only within themselves



Picture courtesy: book by Leskovec, Rajaraman and Ullman

Dead ends

A tiny web



- Let's make C a dead end
- *M* is not stochastic anymore, rather *substochastic*
  - The 3rd column sum = 0 (not 1)
- Now the iteration $v := Mv$ takes all probabilities to zero

M

0	1/2	**0**	0
1/3	0	**0**	1/2
1/3	0	**0**	1/2
1/3	1/2	**0**	0

v

1/4
1/4
1/4
1/4

Mv

3/24
5/24
5/24
5/24

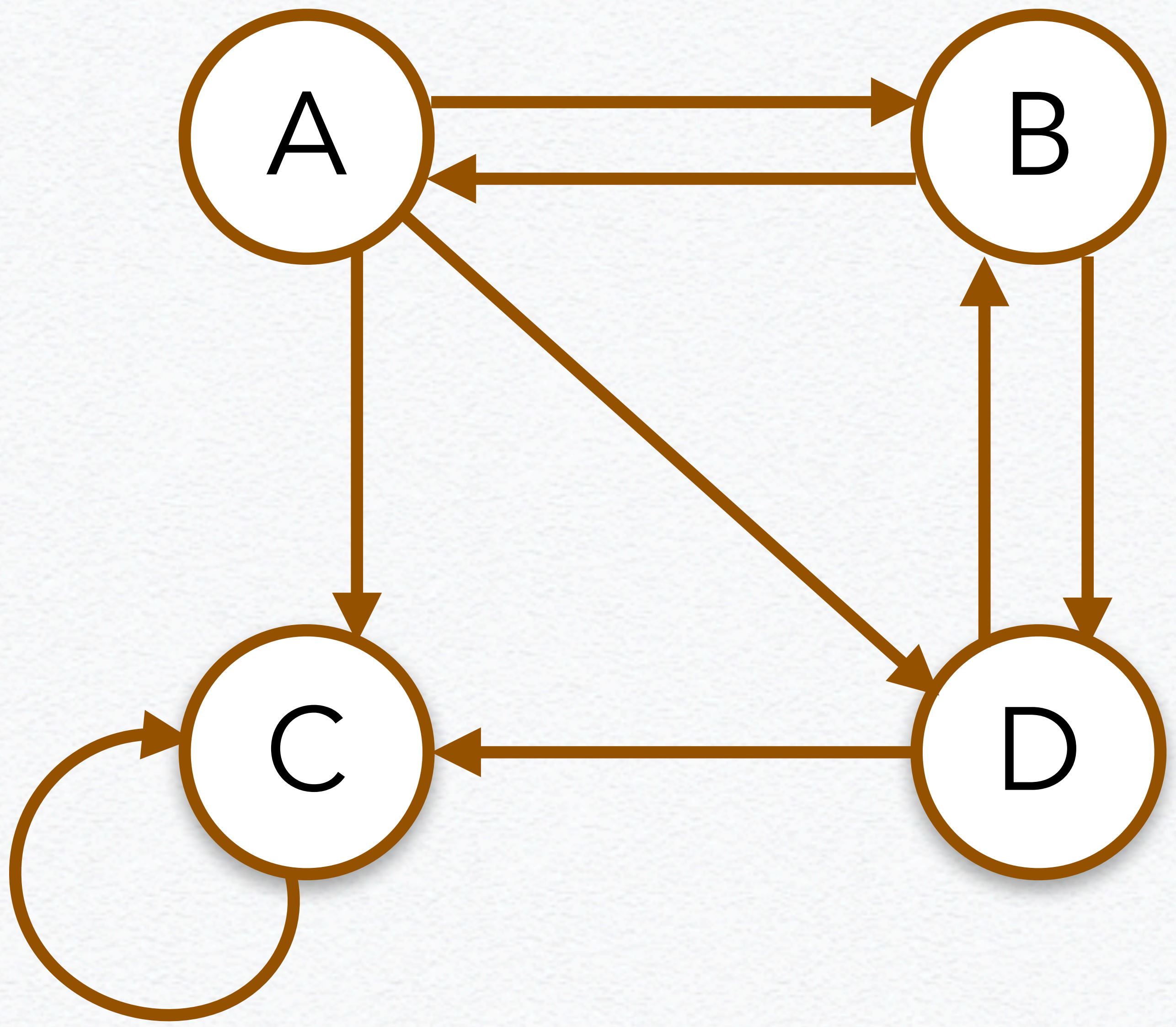
M^2v

5/48
7/48
7/48
7/48

$\cdots \longrightarrow$

0
0
0
0

Spider traps



- Let C be a one node spider trap
- Now the iteration $v := Mv$ takes all probabilities to zero except the spider trap
- The spider trap gets all the PageRank

M

0	1/2	0	0
1/3	0	0	1/2
1/3	0	**1**	1/2
1/3	1/2	0	0

v	Mv	M^2v	$\cdots \longrightarrow$
1/4	3/24	5/48	0
1/4	5/24	7/48	0
1/4	11/24	29/48	1
1/4	5/24	7/48	0

Teleportation

- Approach to handle dead-ends and spider traps
- Teleportation: the surfer may teleport to any other node with some probability
- Idealized PageRank: iterate $v_k = Mv_{k-1}$
- PageRank with teleportation

with probability β continue to an outlink

with probability $(1-\beta)$ teleport (leave and join at another node)

$$v_k = \beta M v_{k-1} + (1-\beta)\frac{\boldsymbol{e}}{n}$$

where β is a constant, usually between 0.8 and 0.9, and $\boldsymbol{e} = (1, \ldots, 1)$

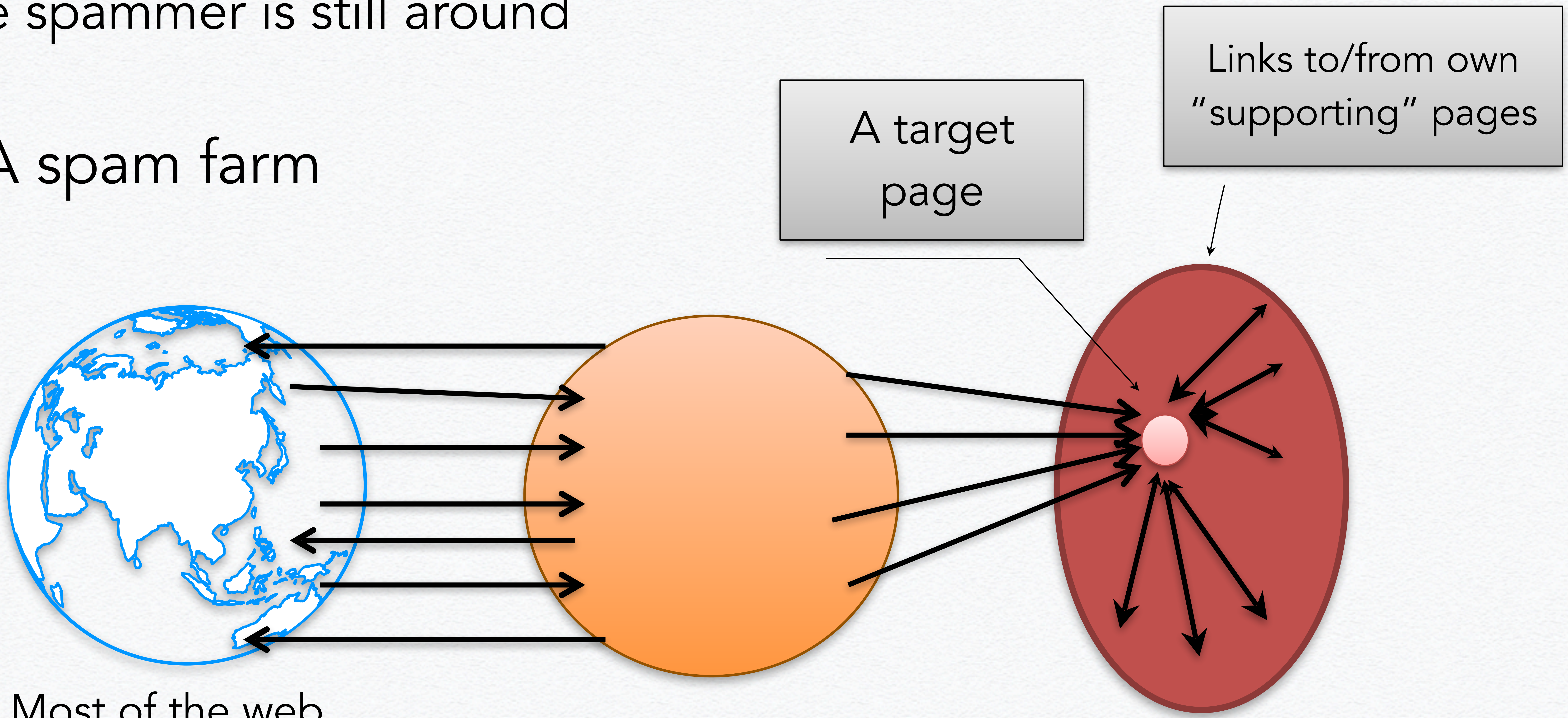
- Intuition: create artificial links to all other pages
- Then, the ergodicity property is also achieved

Link spam

Google took care of the term spam, but …

The spammer is still around

A spam farm



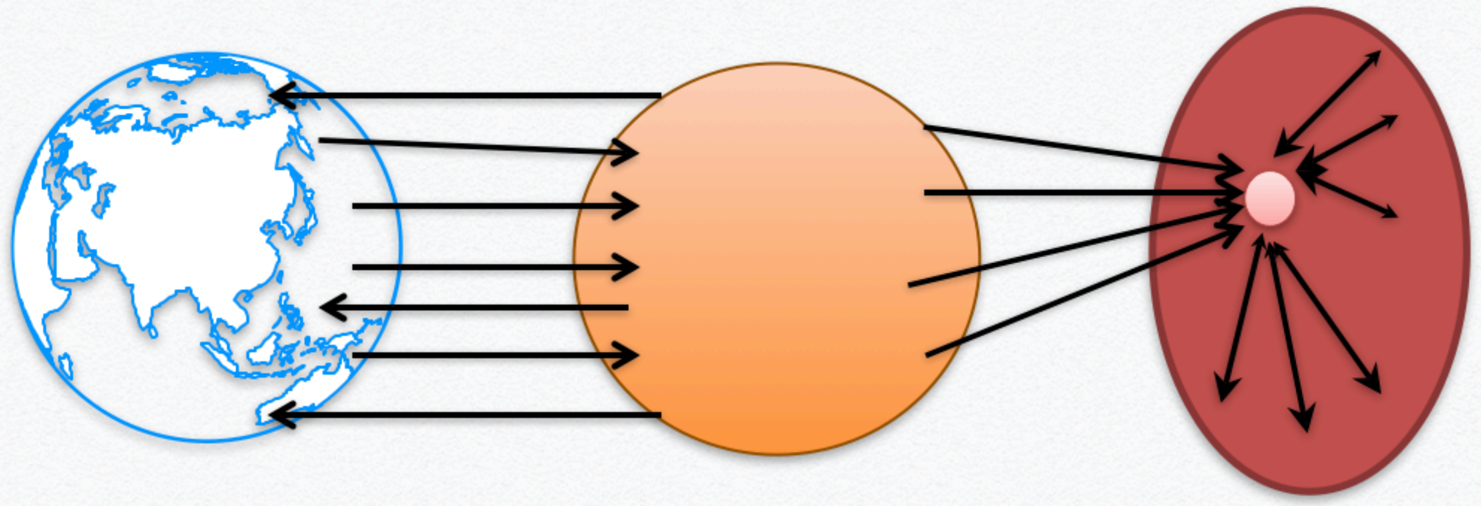
Most of the web
Inaccessible to the spammer

Accessible pages (blogs, sites where spammer can leave comments)

Pages controlled by the spammer (own pages)

Analysis of a spam farm

- Setting
  - Total #of pages in the web = n
  - Target page T, with m supporting pages
  - Let x be the PageRank contributed by accessible pages (sum of all PageRank of accessible pages times β)
  - How much y = PageRank of the target page can be?



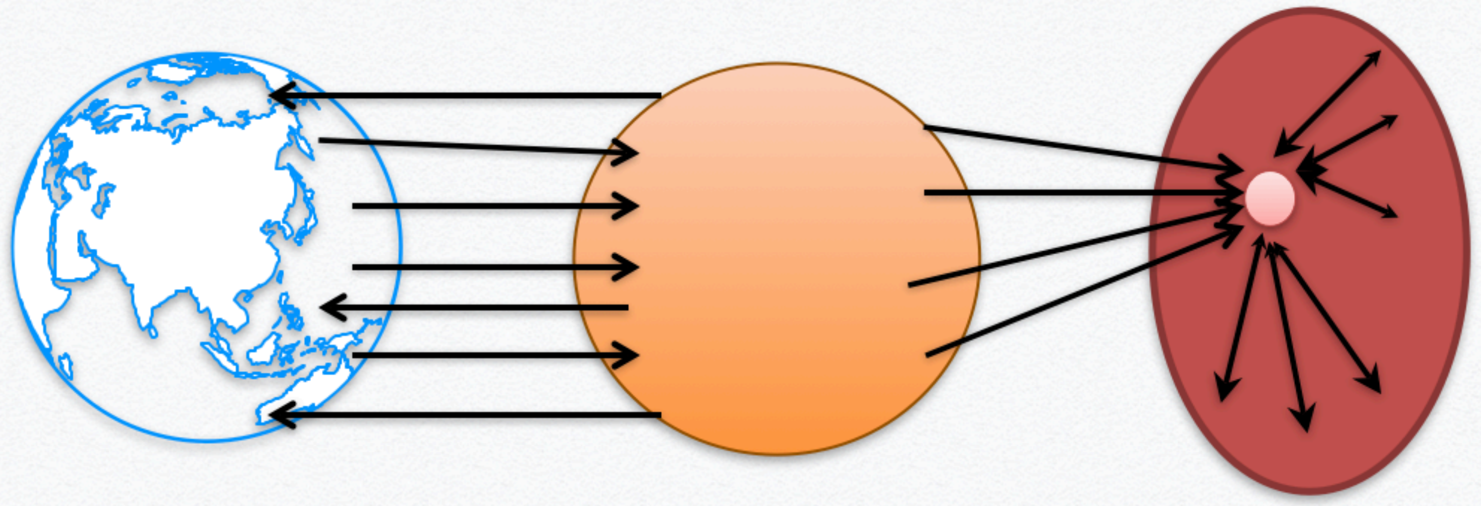
- PageRank of every supporting page

$$\frac{\beta y}{m} + \frac{1-\beta}{n}$$

Contribution from the target page with PageRank y

Share of PageRank among all pages in the web

Analysis of a spam farm (continued)



- Three sources contribute to PageRank
  1. Contribution from accessible pages = x
  2. Contribution from supporting pages = $\beta\left(\frac{\beta y}{m} + \frac{1-\beta}{n}\right)$
  3. The n-th share of the fraction $(1-\beta)/n$ [negligible]
- So, we have $y = x + \beta m\left(\frac{\beta y}{m} + \frac{1-\beta}{n}\right)$
- Solving for y, we get $y = \frac{x}{1-\beta^2} + \frac{\beta}{1+\beta} \times \frac{m}{n}$
- If β = 0.85, then $y = 3.6x + 0.46\frac{m}{n}$
- External contribution up by 3.6 times, plus 46% of the fraction of the PageRank from the web

TrustRank and Personalied PageRank

- The teleportation scheme: $v_k = \beta \times M v_{k-1} + (1-\beta) \times \begin{bmatrix} 1/n \\ \vdots \\ 1/n \end{bmatrix}$

Some nodes can be given higher priority by changing this vector here

- A set S of trustworthy pages where the spammers cannot place links
  - Wikipedia (after moderation), university pages, …
- Compute **TrustRank**: $v_k = \beta M v_{k-1} + (1-\beta)\dfrac{\boldsymbol{e}_S}{|S|}$

  where $\boldsymbol{e}_S$ is the vector with entry = 1 for all pages in S and 0 otherwise
- The random surfers are introduced only at trusted pages
- Spam mass = PageRank – TrustRank
- High spam mass $\Longrightarrow$ likely to be spam
- Similarly, **topic sensitive (personalized)** PageRank
  - For example, prioritize sports pages to create PageRank for users who like sports
  - Combine two or more topic-sensitive PageRanks making it suitable to user profiles

References and further reading

- Leskovec, Rajaraman and Ullman. Mining of Massive Datasets.
- Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1999
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. "Introduction to Information Retrieval", Cambridge University Press. 2008.