Page Rank

Searching in Information Networks

We need a way for filtering the most **important** data

What is important?

An observation

Web search query must return a result very quickly

- We cannot have human experts
- We need an automatic way for evaluating importance

What is important?

An observation

Web search query must return a result very quickly

- We cannot have human experts
- We need an automatic way for evaluating importance

What is importance?

Try to define it. . .

What is important?

An observation

Web search query must return a result very quickly

- We cannot have human experts
- We need an automatic way for evaluating importance

What is importance?

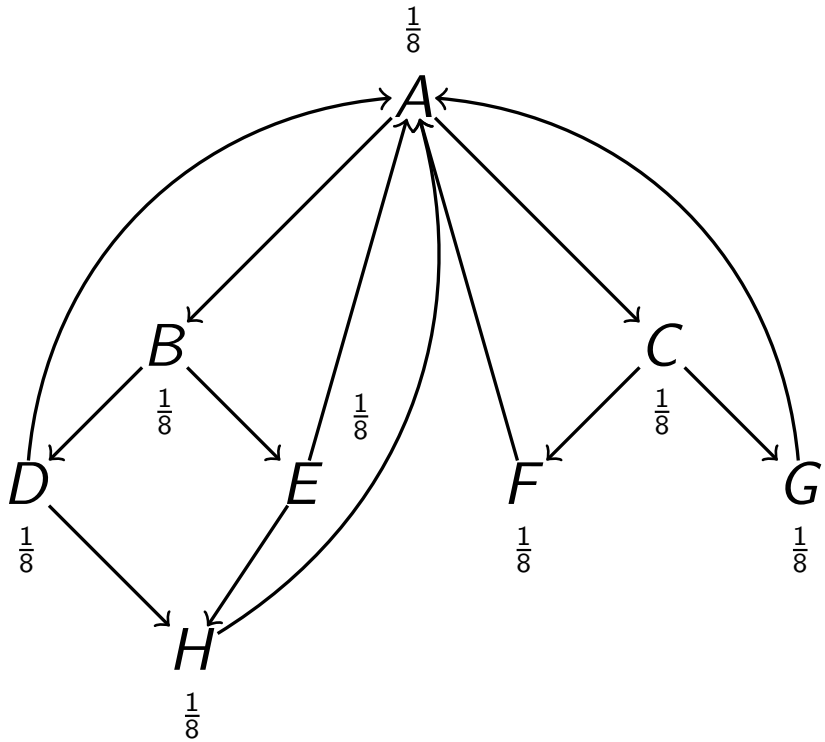
Try to define it. . .

Importance according to PageRank

A web page is important if many important pages are linked to it

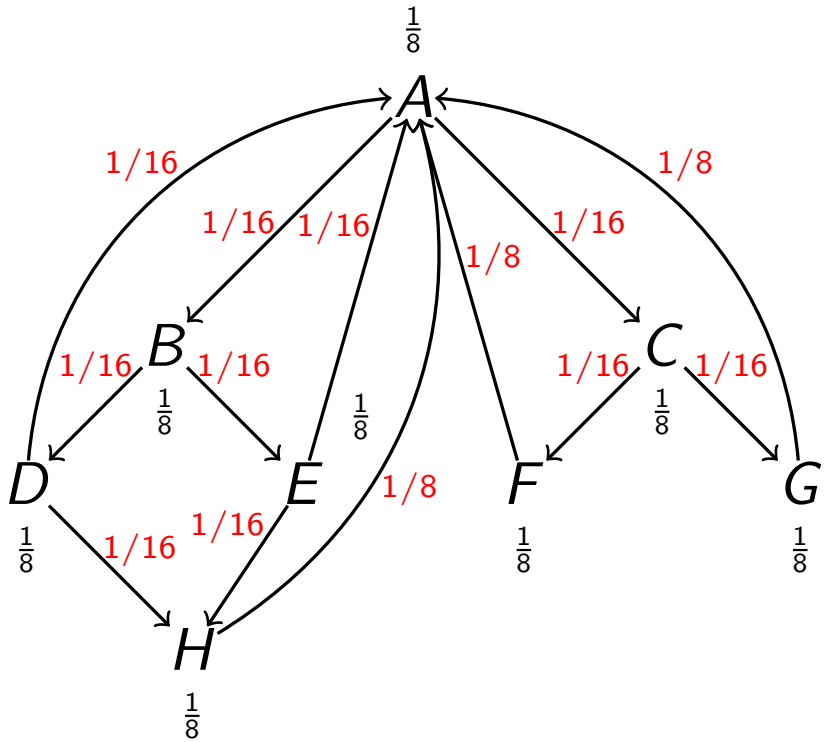
Basic PageRank

1. Start with each page having the same rank $1/n$



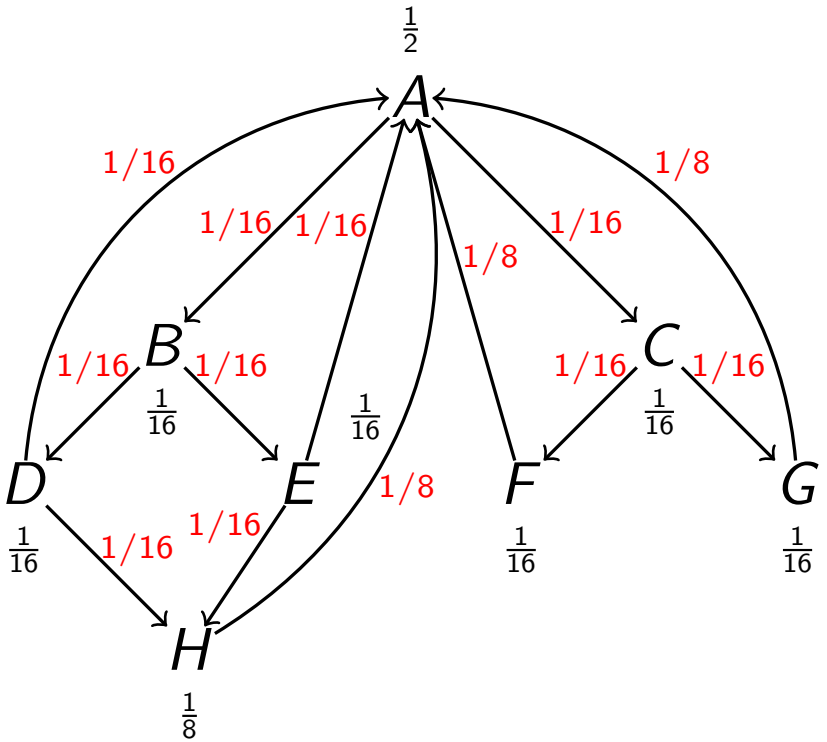
Basic PageRank

1. Start with each page having the same rank $1/n$
2. Each page splits its rank among linked pages



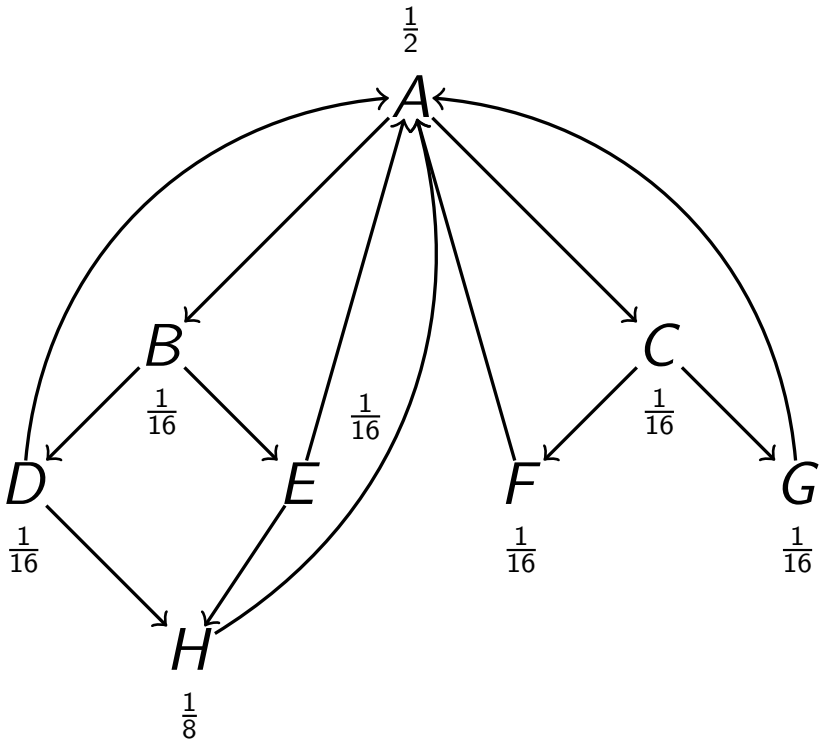
Basic PageRank

1. Start with each page having the same rank $1/n$
2. Each page splits its rank among linked pages
3. Each pages updates its rank as the sum of received shares



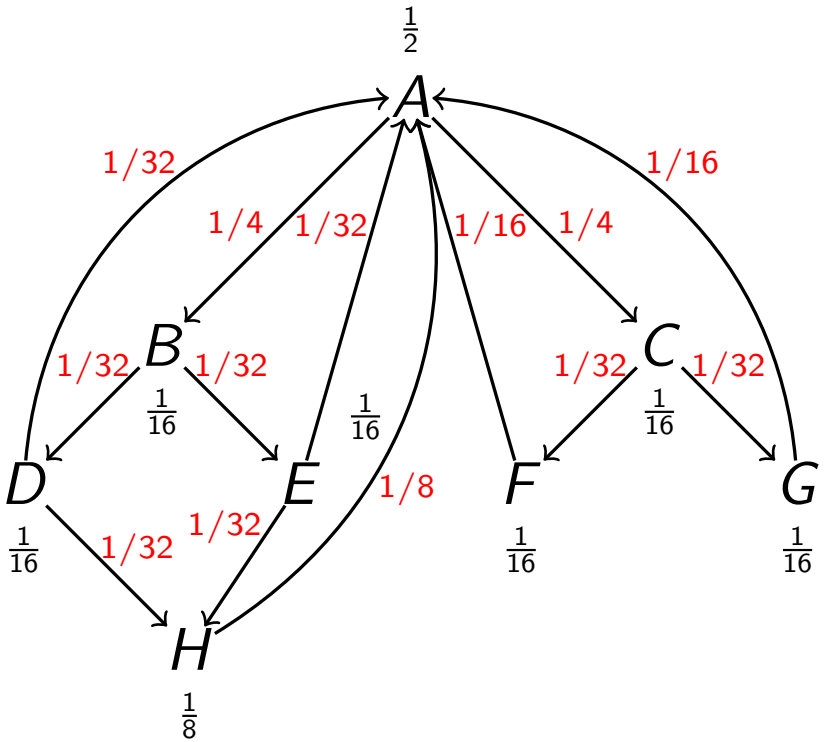
Basic PageRank

1. Start with each page having the same rank $1/n$
2. Each page splits its rank among linked pages
3. Each pages updates its rank as the sum of received shares
4. Repeat from 2



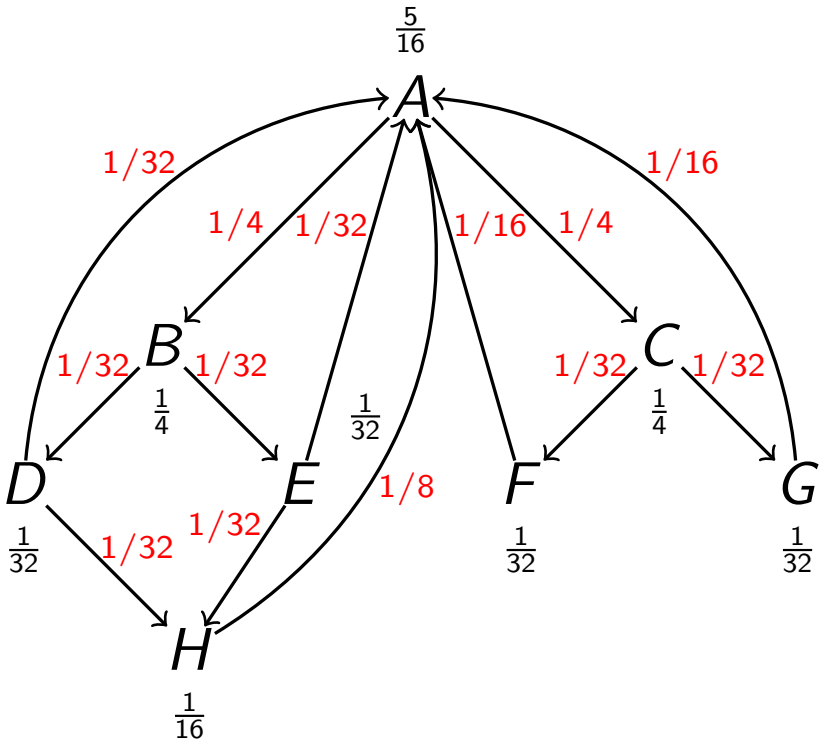
Basic PageRank

1. Start with each page having the same rank $1/n$
2. Each page splits its rank among linked pages
3. Each pages updates its rank as the sum of received shares
4. Repeat from 2



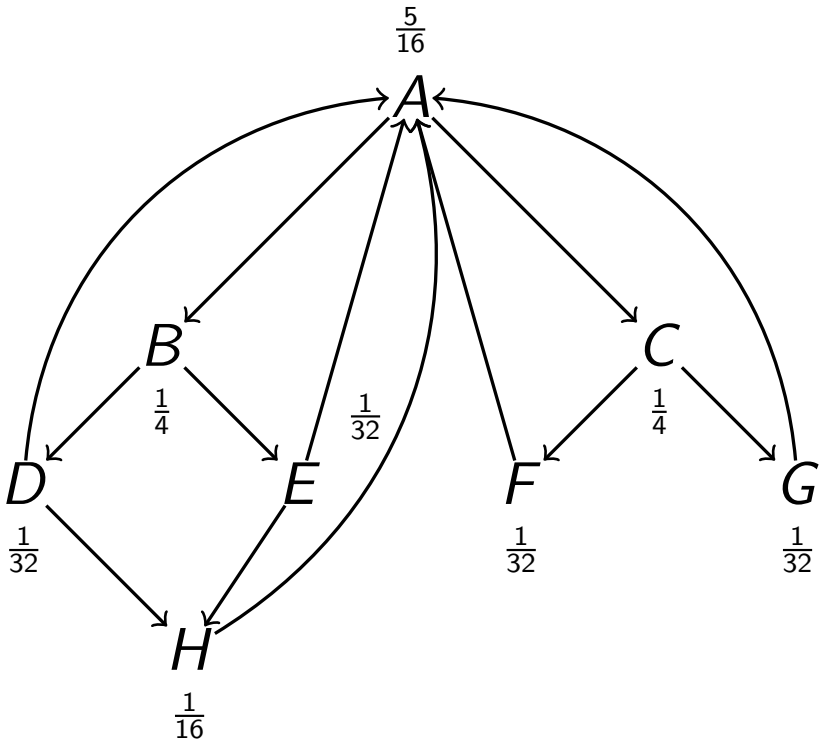
Basic PageRank

1. Start with each page having the same rank $1/n$
2. Each page splits its rank among linked pages
3. Each pages updates its rank as the sum of received shares
4. Repeat from 2



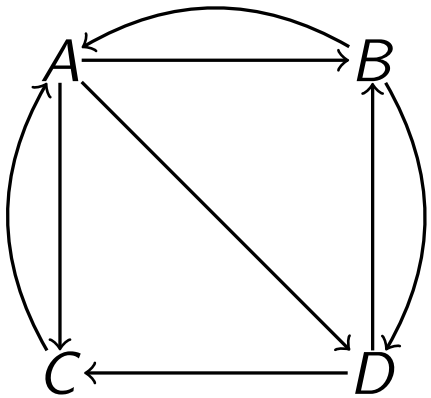
Basic PageRank

1. Start with each page having the same rank $1/n$
2. Each page splits its rank among linked pages
3. Each pages updates its rank as the sum of received shares
4. Repeat from 2



Math enters...

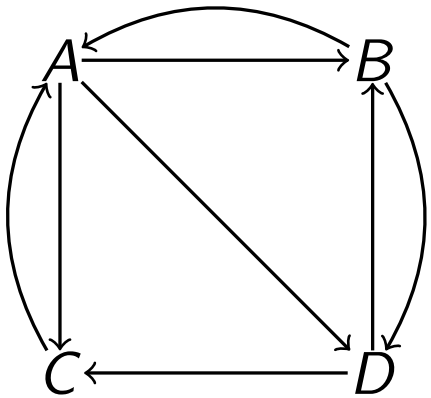
- Transition matrix of the web



$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

Math enters...

- Transition matrix of the web



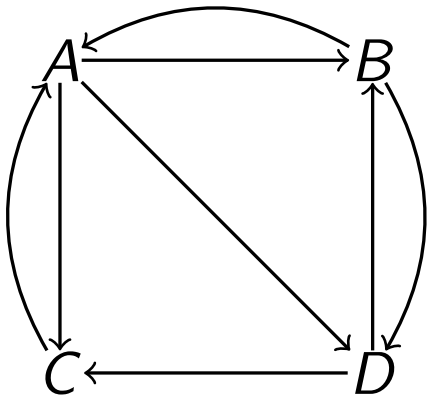
$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

- Rank vector

$$\mathbf{v} = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

Math enters...

- Transition matrix of the web



$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

- Rank vector

$$\mathbf{v} = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

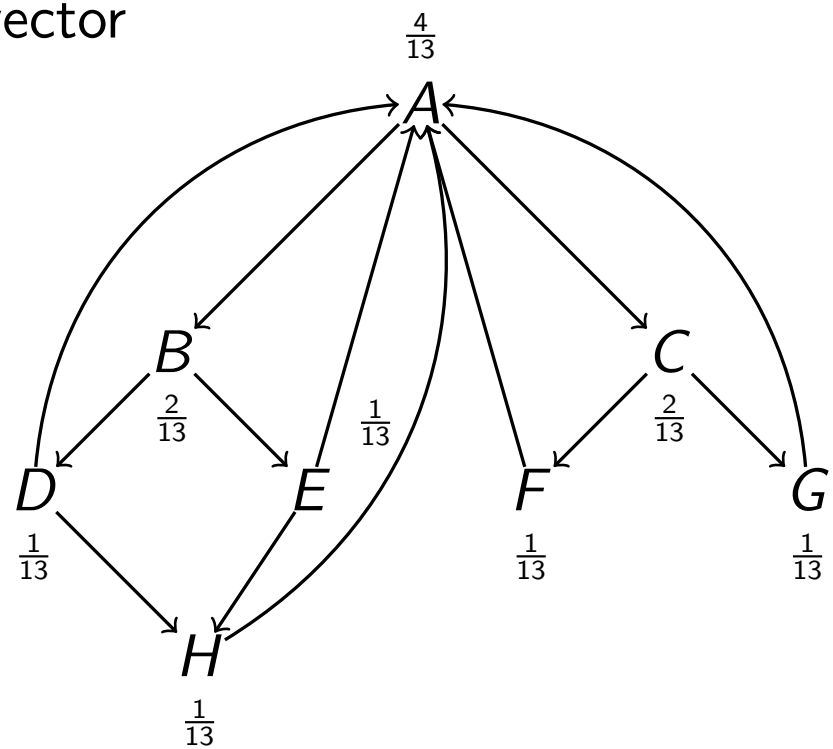
- Update process

$$\mathbf{v}' = M\mathbf{v}$$

Page Rank corresponds to a Markov chain

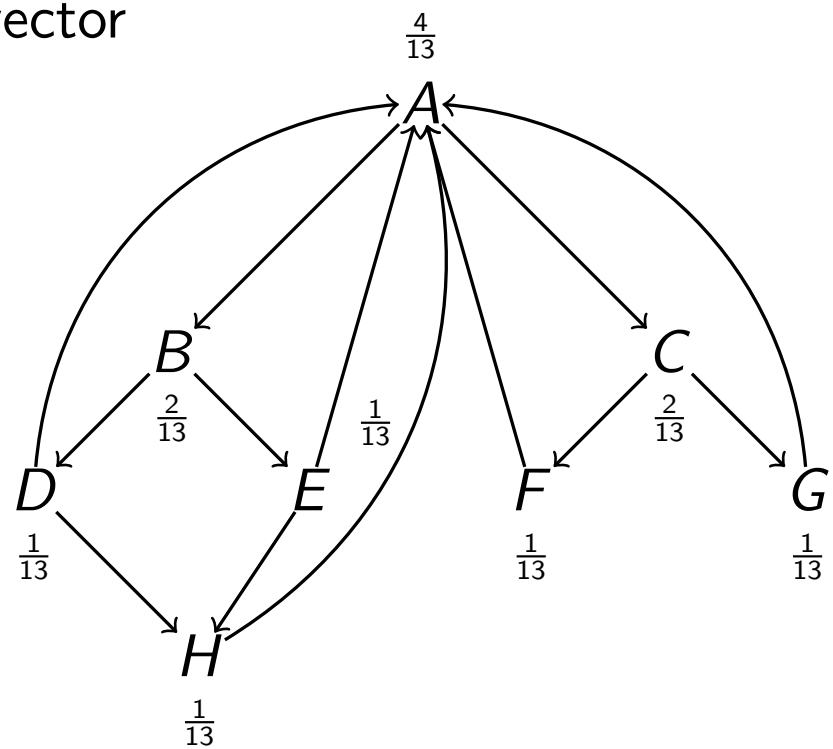
Page Rank corresponds to a Markov chain

- If the graph is strongly connected, then there is a unique limiting rank vector



Page Rank corresponds to a Markov chain

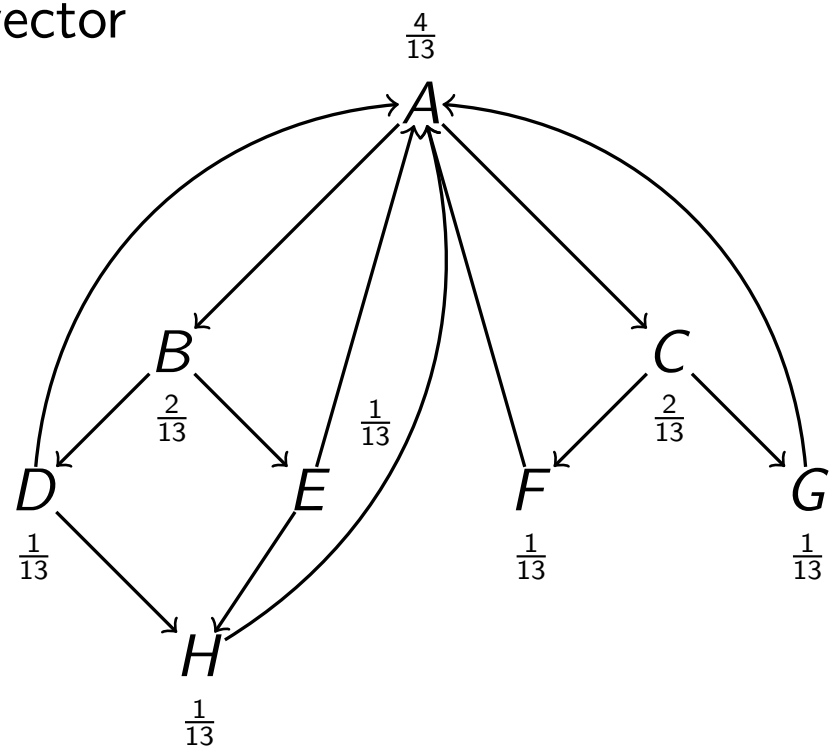
- If the graph is strongly connected, then there is a unique limiting rank vector



  - The rank does not depend on the starting vector

Page Rank corresponds to a Markov chain

- If the graph is strongly connected, then there is a unique limiting rank vector



  - The rank does not depend on the starting vector
  - Experimentally 50-75 repetition are sufficient to converge (within the error limits of double-precision arithmetic)

Page Rank and the experience of a random surfer

An alternative view of Page Rank

- The behavior of a random surfer...
  - Once he lands on a web page...
  - he follows a random link of that page

Page Rank and the experience of a random surfer

An alternative view of Page Rank

- The behavior of a random surfer...
  - Once he lands on a web page...
  - he follows a random link of that page
- This process is called random walk

Page Rank and the experience of a random surfer

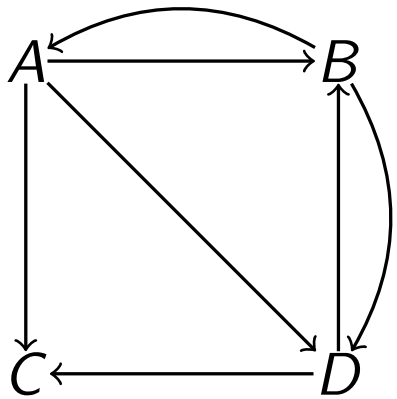
An alternative view of Page Rank

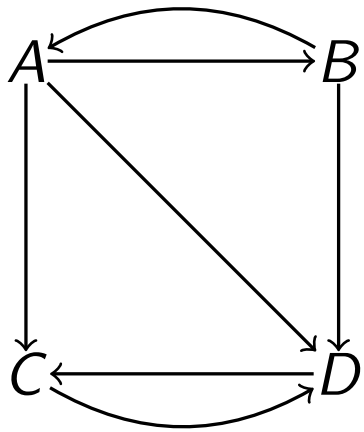
- The behavior of a random surfer...
  - Once he lands on a web page...
  - he follows a random link of that page
- This process is called random walk
- Page rank measures how often the random surfer can be found on a given page

Some problems

The problem

- There are dead ends or spider traps

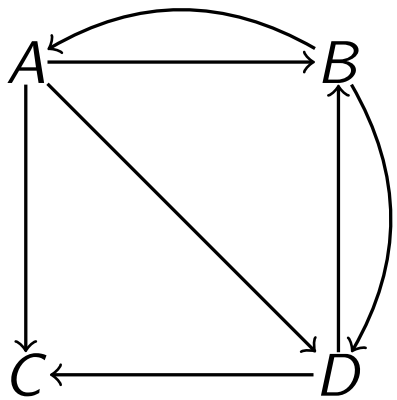


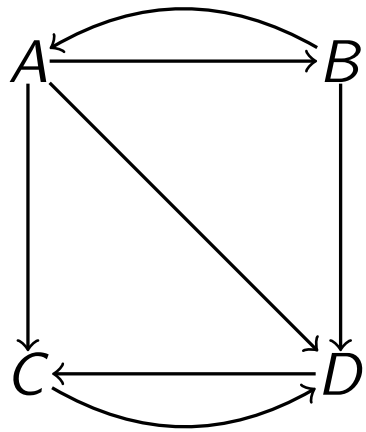


Some problems

The problem

- There are dead ends or spider traps



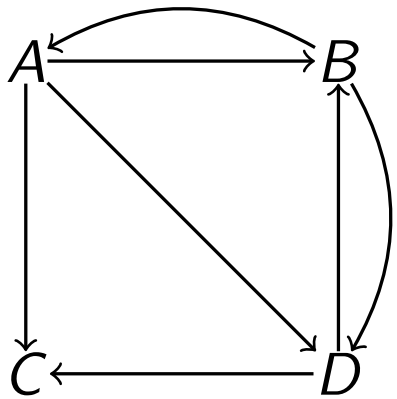


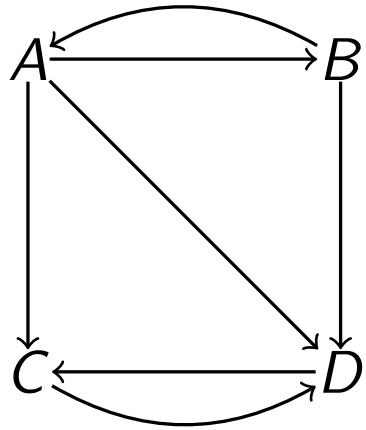
- It is not only a theoretical problem (recall the web structure)

Some problems

The problem

- There are dead ends or spider traps





- It is not only a theoretical problem (recall the web structure)

The solution: The taxation principle

At each step the random surfer has a probability to be teleported to a random page

Page Rank

1. Start with each page having the same rank $1/n$
2. Each page splits a fraction s of its rank among linked pages
3. Each page splits the remaining fraction $1 - s$ among all pages
4. Each pages updates its rank as the sum of received shares
5. Repeat from 2

Page Rank

1. Start with each page having the same rank $1/n$
2. Each page splits a fraction s of its rank among linked pages
3. Each page splits the remaining fraction $1 - s$ among all pages
4. Each pages updates its rank as the sum of received shares
5. Repeat from 2

Observations

- The process is still a Markov chain

Page Rank

1. Start with each page having the same rank $1/n$
2. Each page splits a fraction s of its rank among linked pages
3. Each page splits the remaining fraction $1 - s$ among all pages
4. Each pages updates its rank as the sum of received shares
5. Repeat from 2

Observations

- The process is still a Markov chain
- The graph is strongly connected

Page Rank

1. Start with each page having the same rank $1/n$
2. Each page splits a fraction s of its rank among linked pages
3. Each page splits the remaining fraction $1 - s$ among all pages
4. Each pages updates its rank as the sum of received shares
5. Repeat from 2

Observations

- The process is still a Markov chain
- The graph is strongly connected
  - There is a unique limiting rank vector

Page Rank

1. Start with each page having the same rank $1/n$
2. Each page splits a fraction s of its rank among linked pages
3. Each page splits the remaining fraction $1 - s$ among all pages
4. Each pages updates its rank as the sum of received shares
5. Repeat from 2

Observations

- The process is still a Markov chain
- The graph is strongly connected
  - There is a unique limiting rank vector
  - The rank does not depend on the starting vector

Page Rank

1. Start with each page having the same rank $1/n$
2. Each page splits a fraction s of its rank among linked pages
3. Each page splits the remaining fraction $1 - s$ among all pages
4. Each pages updates its rank as the sum of received shares
5. Repeat from 2

Observations

- The process is still a Markov chain
- The graph is strongly connected
  - There is a unique limiting rank vector
  - The rank does not depend on the starting vector
  - The rank depends on s
    - Usually s is between 0.8 and 0.9
    - It reduces sensitivity to addition or deletion of pages

Alternatives to PageRank: Hubs and authorities

- Authorities: Valuable pages providing infos about a topic
- Hubs: Pages telling where to go for infos about a topic

Alternatives to PageRank: Hubs and authorities

- Authorities: Valuable pages providing infos about a topic
- Hubs: Pages telling where to go for infos about a topic

Who is the best authority? Who is the best suggester?

Alternatives to PageRank: Hubs and authorities

- Authorities: Valuable pages providing infos about a topic
- Hubs: Pages telling where to go for infos about a topic

Who is the best authority? Who is the best suggester?

- Hubs value and authority value

Alternatives to PageRank: Hubs and authorities

- Authorities: Valuable pages providing infos about a topic
- Hubs: Pages telling where to go for infos about a topic

Who is the best authority? Who is the best suggester?

- Hubs value and authority value
- HITS: authorities enforce hubbiness and hubs enforces authoritativeness

Alternatives to PageRank: Hubs and authorities

- Authorities: Valuable pages providing infos about a topic
- Hubs: Pages telling where to go for infos about a topic

Who is the best authority? Who is the best suggester?

- Hubs value and authority value
- HITS: authorities enforce hubbiness and hubs enforces authoritativeness
- It is a Markov chain

Alternatives to PageRank: Hubs and authorities

- Authorities: Valuable pages providing infos about a topic
- Hubs: Pages telling where to go for infos about a topic

Who is the best authority? Who is the best suggester?

- Hubs value and authority value
- HITS: authorities enforce hubbiness and hubs enforces authoritativeness
- It is a Markov chain
  - There is a limiting distribution

Alternatives to PageRank: Hubs and authorities

- Authorities: Valuable pages providing infos about a topic
- Hubs: Pages telling where to go for infos about a topic

Who is the best authority? Who is the best suggester?

- Hubs value and authority value
- HITS: authorities enforce hubbiness and hubs enforces authoritativeness
- It is a Markov chain
  - There is a limiting distribution
- An extension of HITS is used by Google and other search engines (e.g., Ask.com)