Inside PageRank

CHAORAN CHEN

10/23/2017

Overview

- Introduction
- Basic properties
- Stochastic interpretation
- Interaction of communities
- Page promotion
- Page computation
- Conclusions

Introduction

Traditional ranking method:

Fisher's scoring, directly derived from Information Retrieval (IR), employ similarity measures, based on a flat, vector-space model of each page.

Limitations:

- Without considering the topological structure of the Web.
- Web spamming

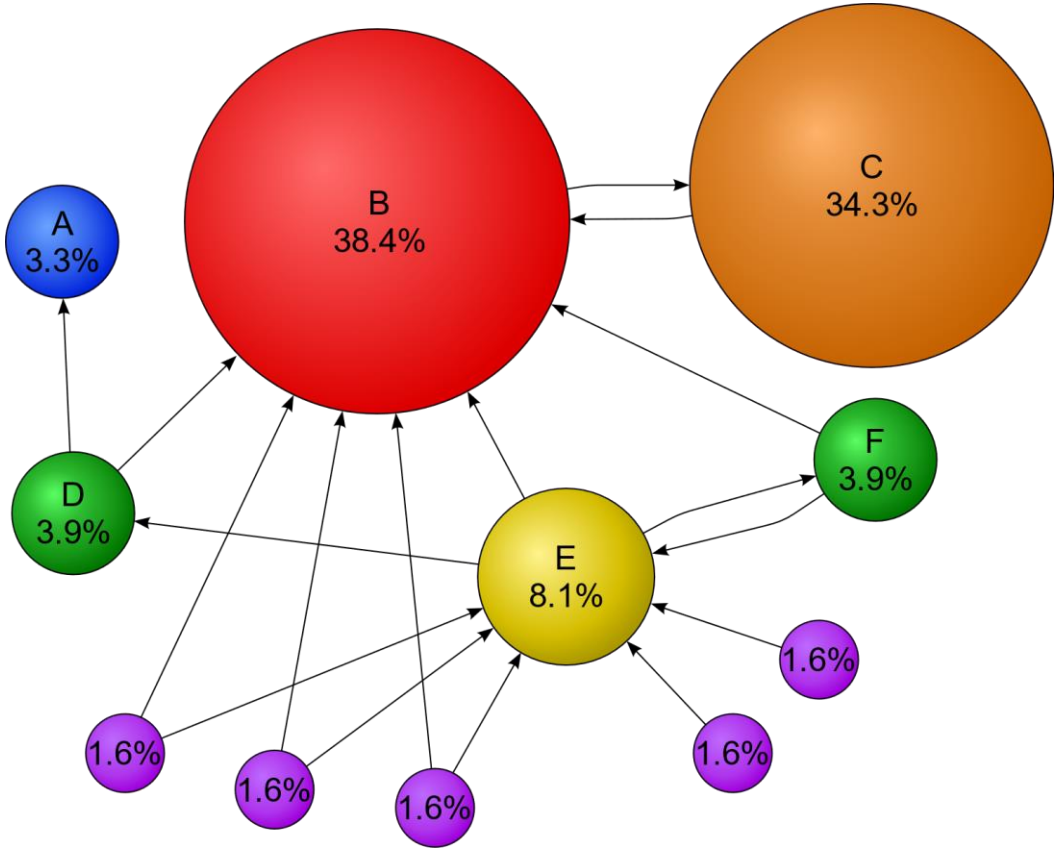
Introduction

Basic assumptions of PageRank:

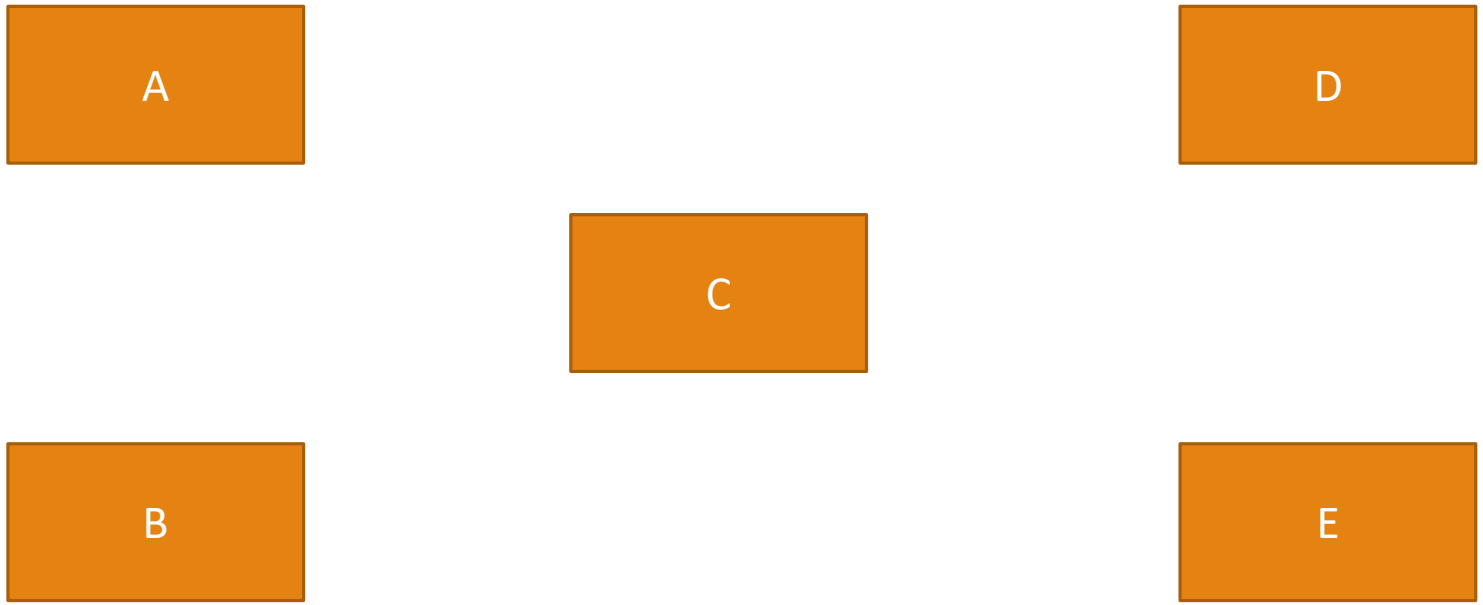
- Quantity: More important websites are likely to receive more links from other websites.
- Quality: Website with higher PageRank will pass higher weight.

Introduction

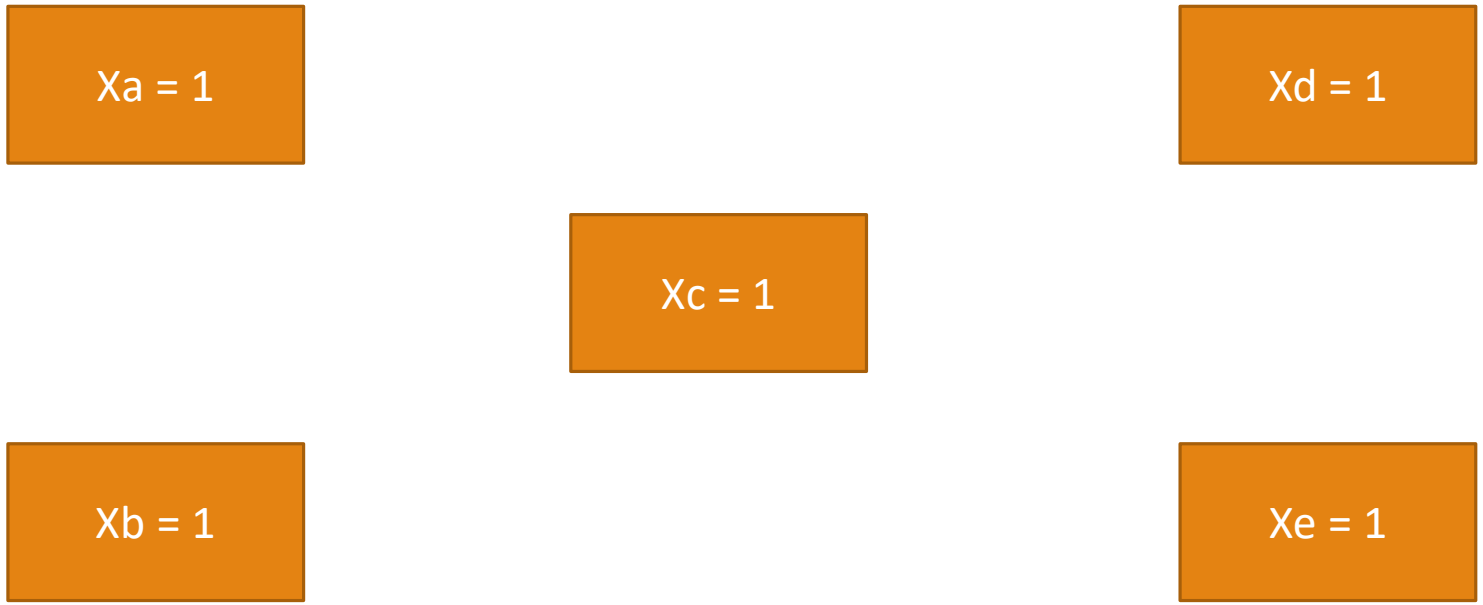
Page C has a higher ranking than Page E.



Introduction

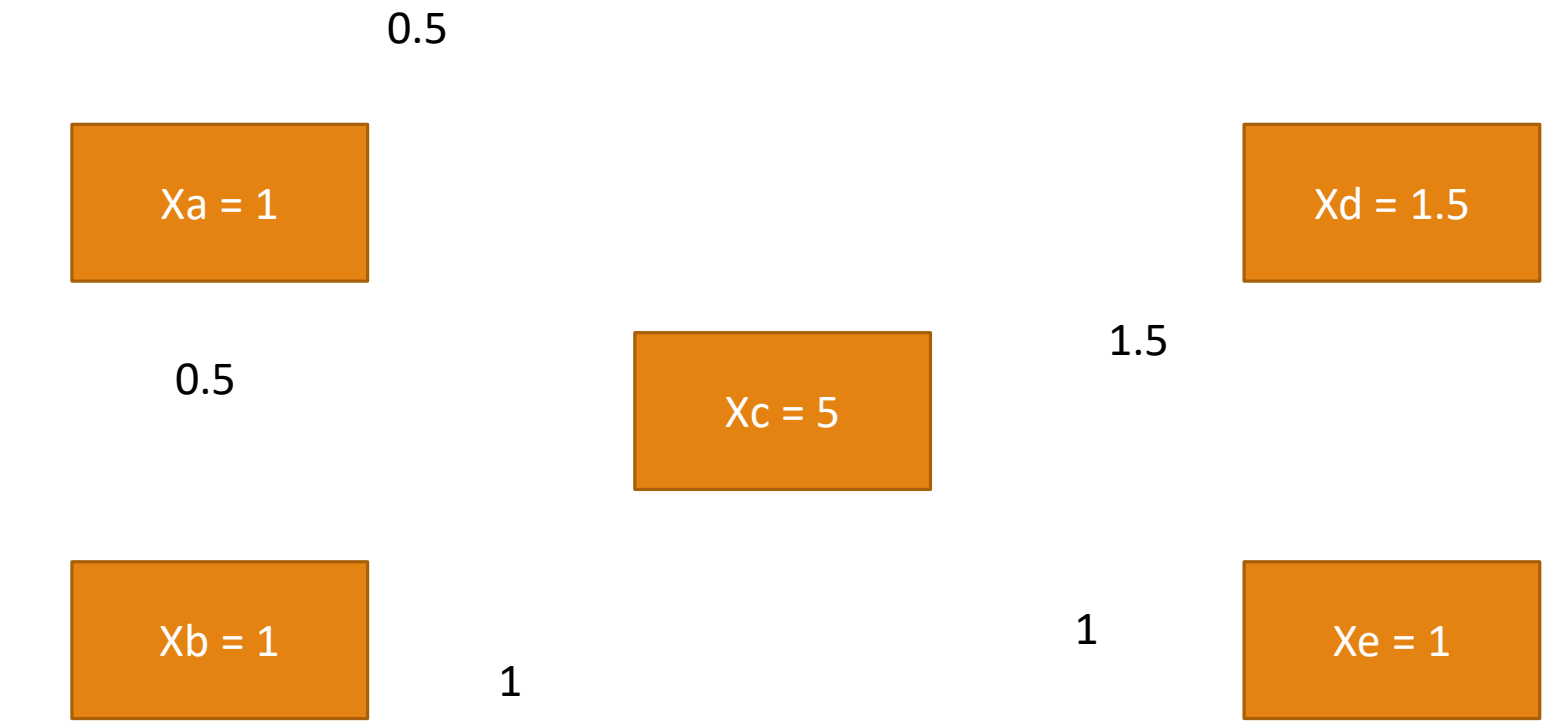


Introduction



Introduction

$$x_p = \sum_{q \in pa[p]} \frac{x_q}{h_q}$$



Ranking: C, D, A, B, E

Introduction

$$x_p = \sum_{q \in pa[p]} \frac{x_q}{h_q}$$

X: ranking

h: outdegree

$$x_p = d \sum_{q \in pa[p]} \frac{x_q}{h_q} + (1 - d)$$

Damping factor d: Make it nearly impossible to deliberately mislead the system in order to get a higher ranking than deserve, d ∈ (0, 1).

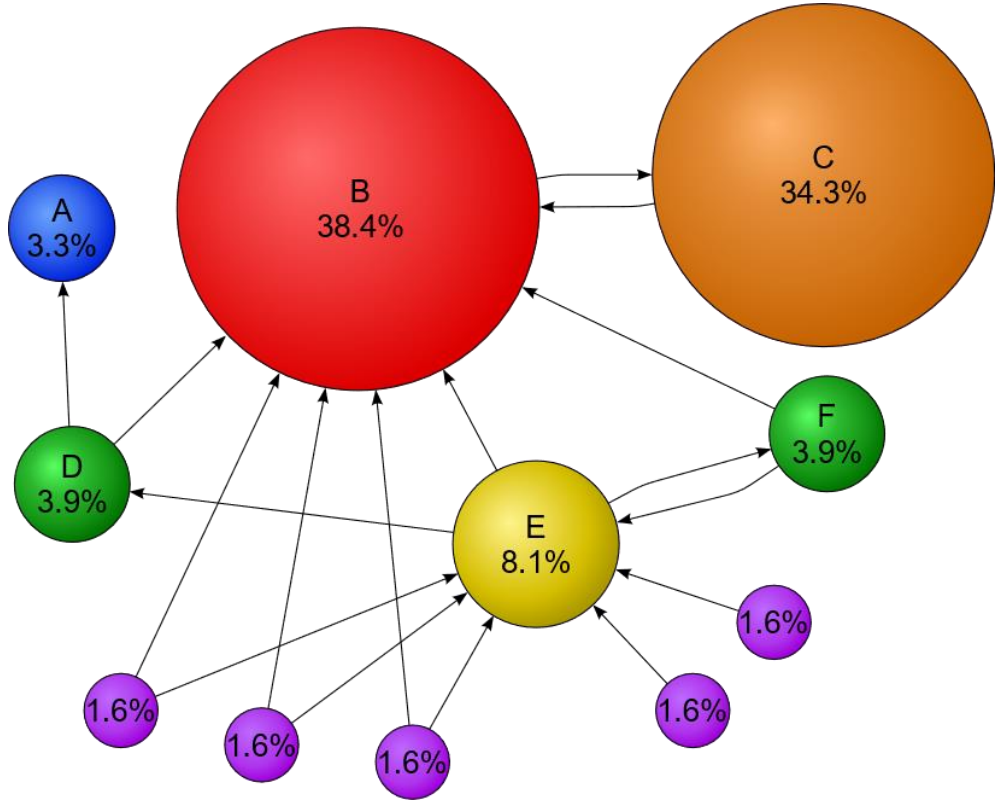
The PageRank theory holds that an imaginary surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the person will continue is a damping factor d.

Dumping factor: close the page and open a new one with probability of (1-d)

Damping factor d = 85% means:

If web surfers who start on a random page have an 85% likelihood of choosing a random link from the page they are currently visiting, and a 15% likelihood of jumping to a page chosen at random from the entire web, they will reach Page E 8.1% of the time. (The 15% likelihood of jumping to an arbitrary page corresponds to a damping factor of 85%.)

Without damping, all web surfers would eventually end up on Pages A, B, or C, and all other pages would have PageRank zero. In the presence of damping, Page A effectively links to all pages in the web, even though it has no outgoing links of its own.



Introduction

From scalar to vector format:

$$x_p = d \sum_{q \in pa[p]} \frac{x_q}{h_q} + (1 - d)$$

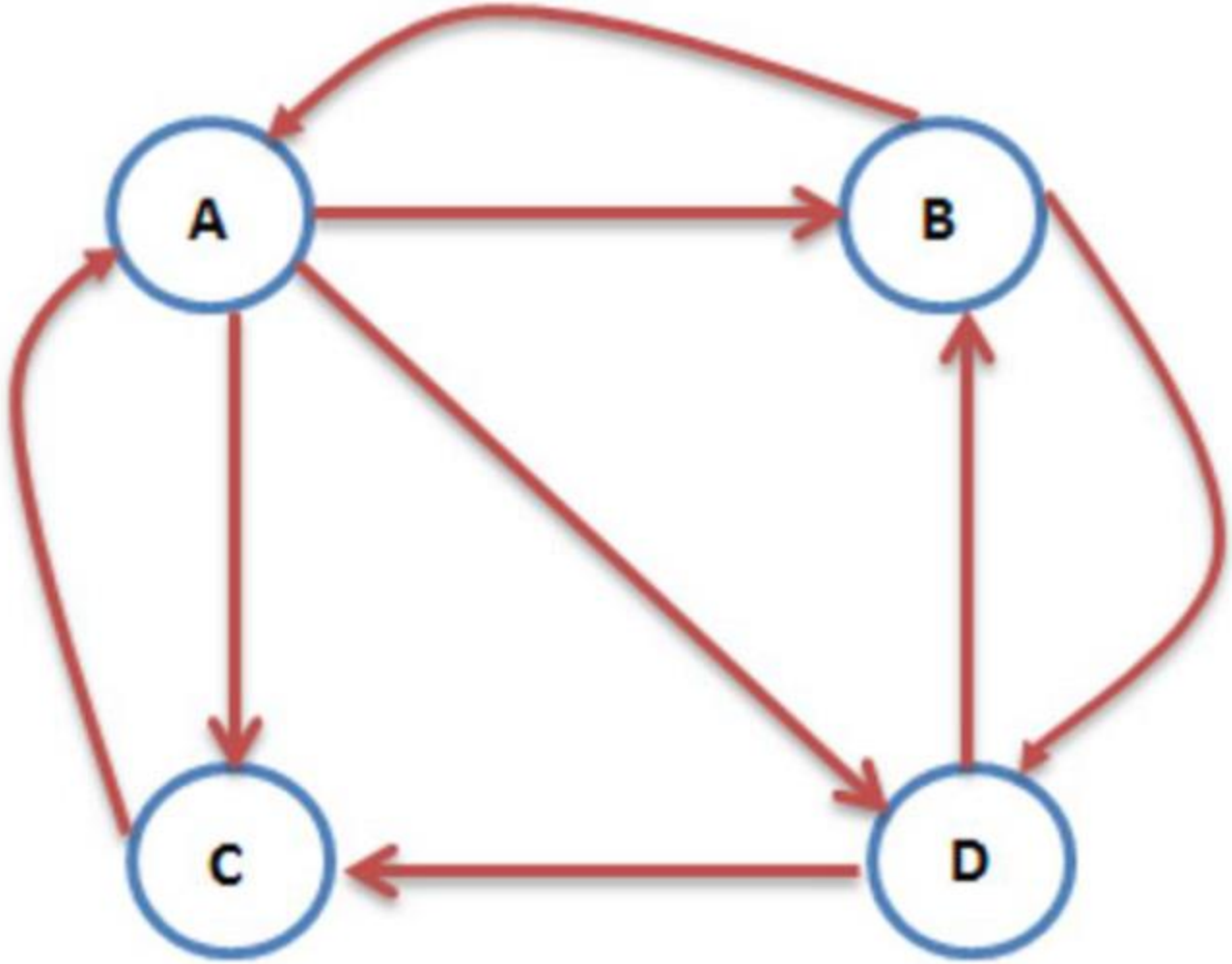
When stacking all x_p into a vector $\boldsymbol{x}$, we get:

$$\boldsymbol{x} = d\boldsymbol{W}\boldsymbol{x} + (1 - d)\boldsymbol{I_N}$$

$\boldsymbol{W}$: Transition matrix, w(i, j) = 1/hj, N x N, each column either sums to 1 or to 0

$\boldsymbol{I_N}$: A column vector with all ones, N x 1

We can prove that it is converged.



A → B, C, D

B → A, D

C → A

D → B, C

A has same jumping probability to B, C, D unless we consider preprocessor after crawling

Transition Matrix

$A \rightarrow B, C, D$

$B \rightarrow A, D$

$C \rightarrow A$

$D \rightarrow B, C$

To\From	WA	WB	WC	WD
WA	0	1/2	1	0
WB	1/3	0	0	1/2
WC	1/3	0	0	1/2
WD	1/3	1/2	0	0

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

PageRank Matrix

Same importance for initial status

	PR0
A	1/4
B	1/4
C	1/4
D	1/4

Initialization

$$x = Wx$$

How to calculate PR1? → PR1 = Transition Matrix * PR0

To\From	WA	WB	WC	WD
WA	0	1/2	1	0
WB	1/3	0	0	1/2
WC	1/3	0	0	1/2
WD	1/3	1/2	0	0

	PR0
A	1/4
B	1/4
C	1/4
D	1/4

→

	PR1
A	9/24
B	5/24
C	5/24
D	5/24

$$x = Wx$$

PR2 = Transition Matrix * PR1

…

PRN = Transition Matrix * PR(N-1)

N is iteration times determined by the user

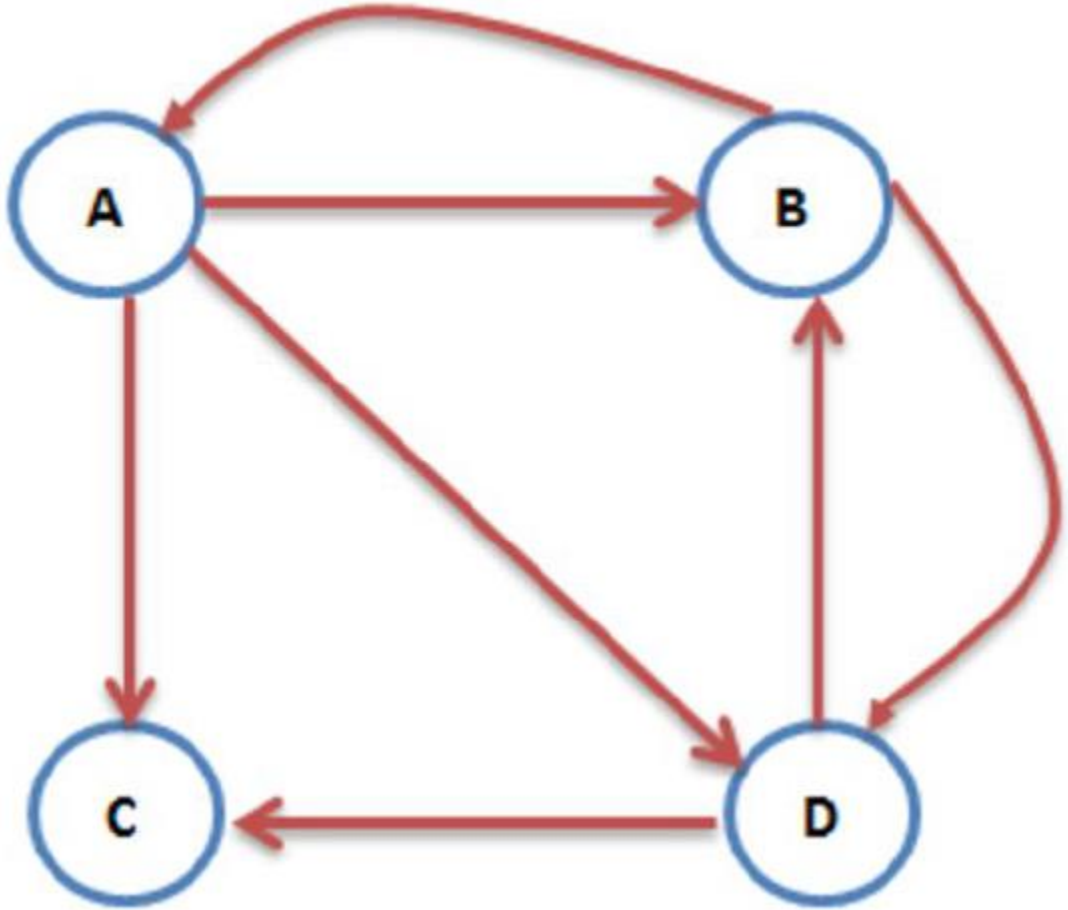
Important pages become closer to 1

Unimportant pages become closer to 0

We can prove the stability:

After N iterations, the matrix PR will finally converge

Basic Properties: dangling pages



To\From	WA	WB	WC	WD
WA	0	1/2	0	0
WB	1/3	0	0	1/2
WC	1/3	0	0	1/2
WD	1/3	1/2	0	0

PR(N) matrix will become zeros finally

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \begin{bmatrix} 3/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix} \begin{bmatrix} 5/48 \\ 7/48 \\ 7/48 \\ 7/48 \end{bmatrix} \begin{bmatrix} 21/288 \\ 31/288 \\ 31/288 \\ 31/288 \end{bmatrix} \cdots \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Basic Properties: removing dangling pages

Dangling pages: The pages that do not contain hyperlinks.

Solution: Introducing a dummy node which has a link to itself.

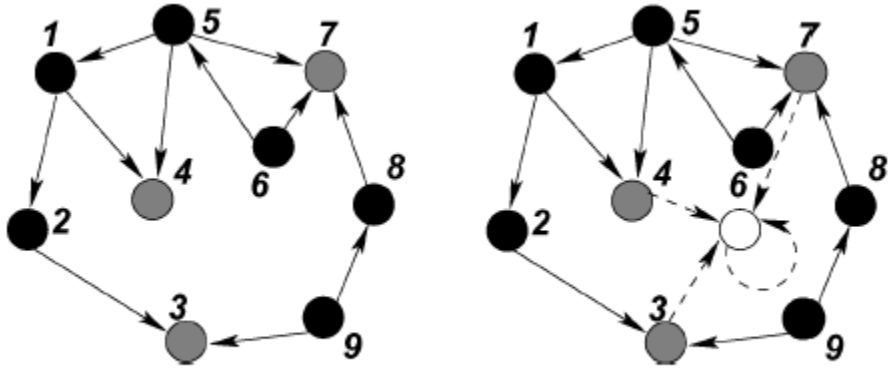


Fig. 4. A trick to eliminate dangling pages: A dummy node with a self-loop is added.

Basic Properties: removing dangling pages

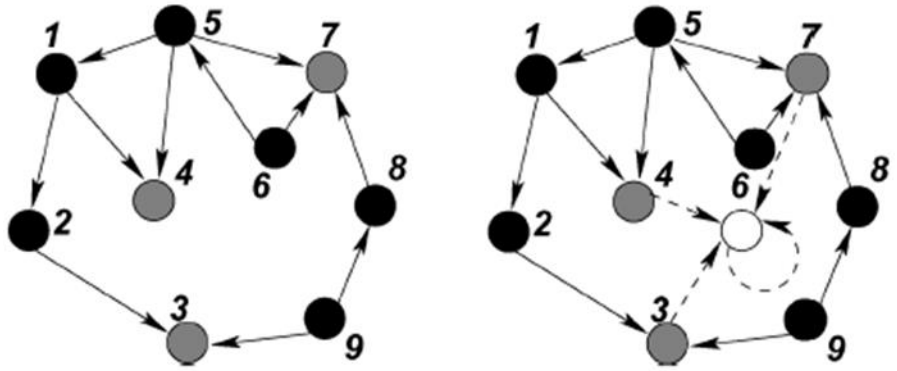


Fig. 4. A trick to eliminate dangling pages: A dummy node with a self-loop is added.

$$\overline{W} = \begin{pmatrix} W & 0 \\ R & 1 \end{pmatrix}$$

$$\overline{W} = \left[\begin{array}{ccccccccc|c} 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 1/2 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right].$$

Proof

$$\boldsymbol{x}(t) = d\boldsymbol{W}\boldsymbol{x}(t-1) + (1-d)\mathbb{1}_N. \tag{3}$$

PROPOSITION 2.1. *Let us consider the dynamical system*

$$\overline{\boldsymbol{x}}(t+1) = d\overline{\boldsymbol{W}}\overline{\boldsymbol{x}}(t) + (1-d)\mathbb{1}_{N+1} \tag{12}$$

related to the extended graph. Then, the following properties hold:

(a) *Eq.* (12) *has a finite stable equilibrium point if and only if Eq.* (3) *has a finite stable equilibrium point.*

(b) *If* $\overline{\boldsymbol{x}}^*$ *is an equilibrium point of* (12) *and* $\boldsymbol{x}^*$ *is an equilibrium point of* (3), *then* $\overline{\boldsymbol{x}}^* = [\boldsymbol{x}^{*\prime}, 1 + \frac{d}{1-d}\boldsymbol{R}\boldsymbol{x}^*]'$.

(c) *If* $\boldsymbol{x}(0) = \mathbb{1}_N$ *and* $\overline{\boldsymbol{x}}(0) = \mathbb{1}_{N+1}$, *then* $\overline{\boldsymbol{x}}(t) = [\boldsymbol{x}(t)', 1 + \sum_{s=0}^{t-1} d^{t-s}\boldsymbol{R}\boldsymbol{x}(s)]'$.

Basic Properties: stability and dynamical updating

PROPOSITION 2.3. *For any Web graph* $\mathbf{G}_W$, *the spectral radius* $\rho(\mathbf{W})$ *fulfills* $\rho(\mathbf{W}) \leq 1$. *Moreover, if* $\mathbf{G}_W$ *has no dangling page, then* $\rho(\mathbf{W}) = 1$ *and* $\mathbb{1}_N$ *is the left eigenvector associated with the largest eigenvalue* $\lambda_{max} = 1$.

Guarantees that the PageRank scheme is well founded whenever $0 \leq d < 1$.

Basic Properties: stability and dynamical updating

LEMMA 2.1. *Let us consider the dynamical system* (3) *and its stationary point* $\boldsymbol{x}^*$. *Let* S *be the set of dangling pages. Then,* $\boldsymbol{x}^* > 0$ *and*

$$\|\boldsymbol{x}^*\|_1 = N - \frac{d}{1-d}\sum_{i\in S} x_i^*. \tag{14}$$

Moreover, if $\boldsymbol{x}(0) = \mathbb{1}_N$, *then, for each* t, $\boldsymbol{x}(t) > 0$ *and*

$$\|\boldsymbol{x}(t)\|_1 = N - \sum_{i\in S}\sum_{k=1}^{t} d^k x_i(t-k)\,. \tag{15}$$

COROLLARY 2.1. *If* $\boldsymbol{G}_W$ *has no dangling page and* $\boldsymbol{x}(0) = \mathbb{1}_N$, *then*

$$\forall\, t \geq 0:\ \|\boldsymbol{x}(t)\|_1 = \|\boldsymbol{x}^*\|_1 = N.$$

Lemma 2.1 also points out that, due to the presence of dangling pages, the Web loses part of its energy. The *energy loss* is represented by the negative term of Eq. (14), $E^{dp} = d/(1-d)\sum_{i\in S} x_i^*$. In the worst (ideal) case, when all the pages are dangling pages, $x_p = (1-d)$ for each p, and, therefore, the loss of energy is dN. If d approaches 1 (e.g., $d = 0.85$ is the value suggested in Brin et al. [1999]), the loss can be an important percentage of the available energy.

Basic Properties: The Dumping Factor Boundary Values

$$x_p = d \sum_{q \in pa[p]} \frac{x_q}{h_q} + (1 - d) .$$

PageRank is strongly affected by d!

If d = 0, all the PageRanks equals to 1.

If d = 1, the sequence might not converge. Many pages would have a zero PageRank.

Basic Properties: The Dumping Factor Boundary Values

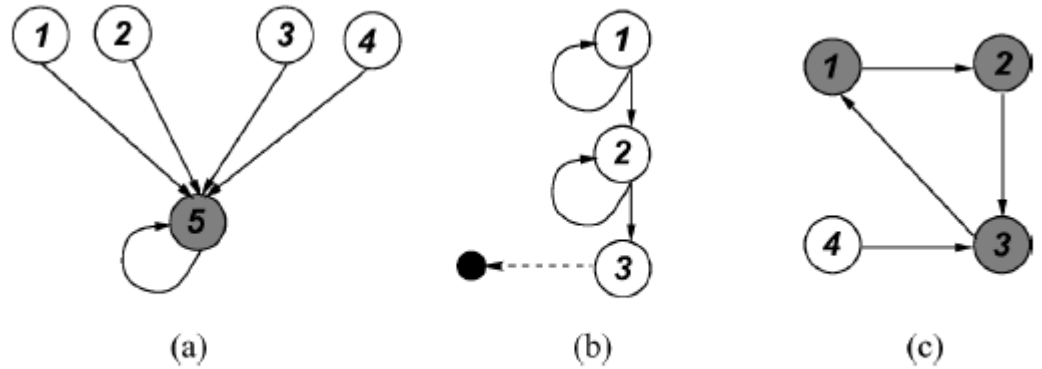


Fig. 5. Examples of essential (grey circles) and inessential nodes (white circles). The black node (connected with a dashed line in (b)) is the dummy node of the extended Web.

A page is **essential** if it belongs to a subgraph where a surfer can be entrapped, that is, a group of connected pages without any way of escaping.

Since the theory of Markov chains can be applied only to stochastic matrices, in order to define the inessential pages, we use the extended Web, which has no dangling pages.

Stochastic Interpretation

Assumption: a random surfer spends a lot of time in important pages: in fact, P(p) describes how often a page will be visited, a probability.(Random Walk theory)

Definition 3.1. Let us denote $\mathcal{S}_i(t) = (p_i(t), a_i(t))$, $p_i(t) \in P \cup \{\text{“}idle\text{”}\}$, $a_i(t) \in \mathcal{A} = \{l, s\}$, and $i = (t_{i(0)}, p_{i(0)})$. $\mathcal{S}_i$ defines a random Markovian *Web surfer,* which starts its navigation from page $p_{i(0)}$ at time $t_{i(0)}$. Function $\mathcal{S}_i$ describes the position on $\boldsymbol{G_W}$ at time $t \geq t_{i(0)}$ and the action which will be performed next. In particular, $\mathcal{S}_i$ can perform two actions $a_i(t) \in \mathcal{A}$:

—follow a hyperlink (action l);

—stop surfing and become “*idle*” (action s).

Stochastic Interpretation

Therefore, we define:

—$\mathcal{P}(a_i(t) = a \mid p_i(t) = p)$ as the probability of performing $a \in \mathcal{A}$, provided that $\mathcal{S}_i$ stays in p at time t;

—$\mathcal{P}(p_i(t+1) = p | p_i(t) = q, a_i(t) = l)$ as the probability of moving from page q to page p, provided that $\mathcal{S}_i$ is located in q at time t and it will follow a hyperlink;

—$\mathcal{P}(p_i(t+1) = p | p_i(t) = q) = \sum_{a \in \mathcal{A}} \mathcal{P}(p_i(t+1) = p \mid p_i(t) = q, a_i(t) = a) \mathcal{P}(a_i(t) = a \mid p_i(t) = q)$ as the probability of moving to page p, provided that $\mathcal{S}_i$ is located in q at time t and the performed action is unknown;

—$\mathcal{P}(p_i(t) = p)$ as the probability that $\mathcal{S}_i$ is located in p at time t.

The expected number of surfers E(np(t + 1)) which lie in p at time t + 1 can be calculated as

$$E(n_p(t+1)) = \sum_{i \in Q} \mathcal{P}(p_i(t+1) = p)$$
$$= \alpha + \sum_{i \in Q} \sum_{q \in P} \mathcal{P}(p_i(t+1) = p \mid p_i(t) = q) E(n_q(t)).$$

Becomes $x_p = d \sum_{q \in pa[p]} \frac{x_q}{h_q} + (1 - d)$ when, for each p, q in P, making the following assumptions:

(1) $\alpha = 1 - d$, that is, $1 - d$ surfers start navigating from each page, at each time step;

Dumping factor

(2) if p is not a dangling page, $\mathcal{P}(a_i(t) = l \mid p_i(t) = p) = d$, that is, every $\mathcal{S}_i$ has the same constant probability d of following a link in each page (and a constant probability $1 - d$ of stopping the navigation);

(3) $\mathcal{P}(p_i(t+1) = p \mid p_i(t) = q, a_i(t) = l) = 1/h_q$ if there exists a link (q, p), and 0 otherwise. Here h_q is the outdegree of page q, that is, the number of outcoming links from q. This assumption makes the surfer "random", since all the outcoming links from a page are followed with the same probability;

Same probability

(4) if p is a dangling page, $\mathcal{P}(a_i(t) = s \mid p_i(t) = p) = 1$, that is, the surfers stop the navigation on dangling pages.

Explain why dp loss of energy

Communities and Energy Balance

A **community** could be a set of pages on a given topic, the researchers' home pages or a Website; the corresponding energy is a measure of its authority.

$$E_I = |I| + E_I^{in} - E_I^{out} - E_I^{dp}.$$

|I|: # of pages, "default energy"
Ein: Page Rank inside the community, communities with many references have a high authority
Eout: Page Rank ouside the community , having hyperlinks outside the community leads to decrease energy
Edp: the presence of pages without hyperlinks yields a loss of energy

Communities and Energy Balance

$$E_I = |I| + E_I^{in} - E_I^{out} - E_I^{dp}.$$

$$E_I \leq |I| + E_I^{in}.$$

The energy of a community is bounded.
Small communities with few references cannot have pages with high score.

Communities and Energy Balance

$$E_I^{in} = \frac{d}{1-d} \sum_{i \in \text{in}(I)} f_i x_i^*,$$

$$E_I^{out} = \frac{d}{1-d} \sum_{i \in \text{out}(I)} (1 - f_i) x_i^*,$$

$$E_I^{dp} = \frac{d}{1-d} \sum_{i \in dp(I)} x_i^*,$$

fi: fraction of the hyperlinks of page i that point to pages in community I with respect to the total number of hyperlinks outgoing from i.

Communities and Energy Balance

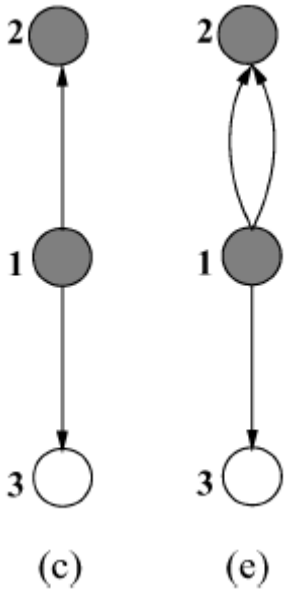
THEOREM 4.2. *Given a community $\boldsymbol{G}_I$, let f_p be the fraction of the hyperlinks of page p that point to pages in $\boldsymbol{G}_I$ with respect to the total number of hyperlinks outgoing from p. Let E_I^{in}, E_I^{out}, and E_I^{dp} be defined by*

$$E_I^{in} = \frac{d}{1-d} \sum_{i \in \text{in}(I)} f_i x_i^*, \quad E_I^{out} = \frac{d}{1-d} \sum_{i \in \text{out}(I)} (1 - f_i) x_i^*, \quad E_I^{dp} = \frac{d}{1-d} \sum_{i \in dp(I)} x_i^*.$$

Then, PageRank $\boldsymbol{x}_I^*$ of $\boldsymbol{G}_I$ satisfies

$$E_I = |I| - E_I^{dp} + E_I^{in} - E_I^{out}. \tag{23}$$

Communities and Energy Balance



f1: 1/2 vs 2/3
Energy lost: 1/2 vs. 1/3

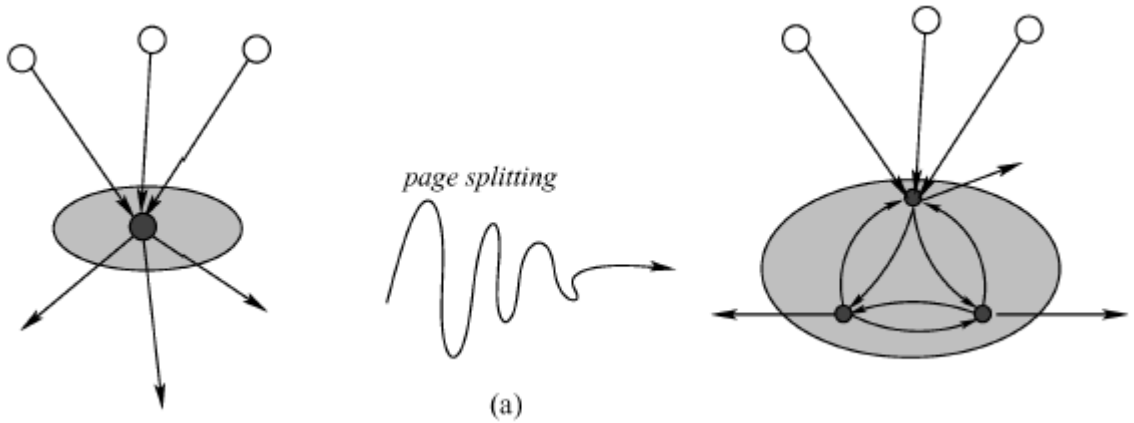
An appropriate organization of a community must avoid energy loss and give rise to a useful distribution of the available energy among the pages.

Page Promotion

Web visibility can be promoted working on both page content and pattern of connections.

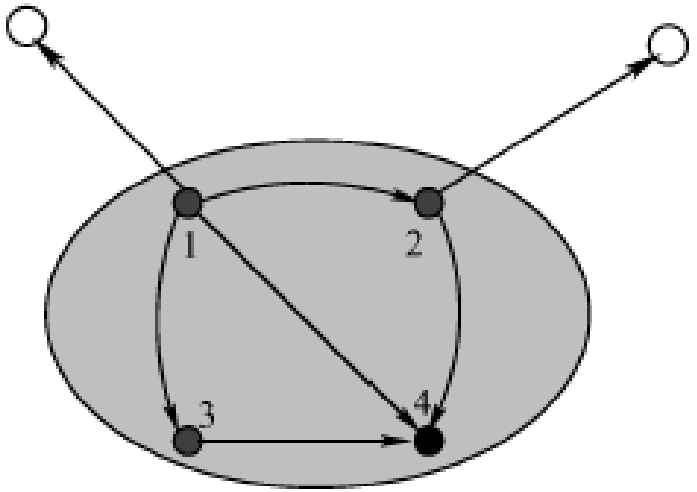
- Energy balance equations derive the following rules which only take into account the topological structure of the community:

1. The same content divided into many small pages yields a higher score than the same content into a single large page.



Page Promotion

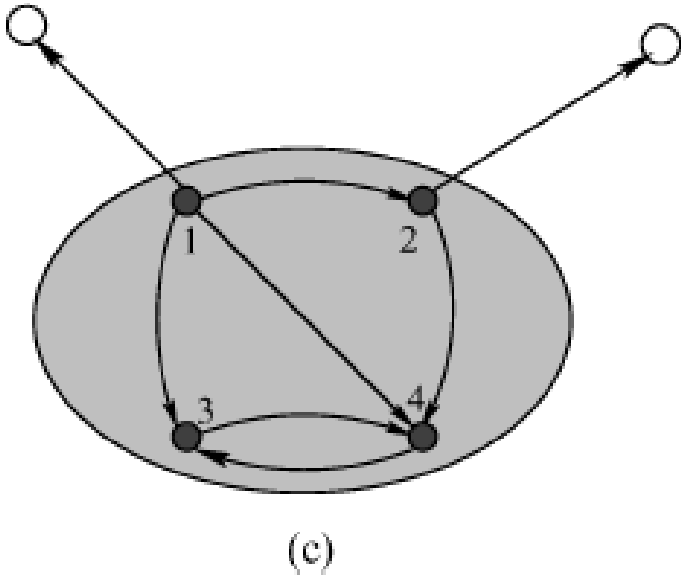
2. Dangling pages give rise to a loss of energy in the community they belong to. The lost energy is small provided that the pages that point to dangling pages have a small score and many hyperlinks pointing to pages of the community.



(b)

Page Promotion

3. Hyperlinks that point outside the community originate a loss of energy, which is high when the hyperlinks belong to pages with high PageRank. The lost energy depends also on the fraction of all the links which point outside the community. Hence, this energy is small whenever the pages pointing outside have many hyperlinks to pages of the community.



(c)

Page Promotion

Instead of changing the topological structure, we can also:

- Exploit external links coming from another community.

PageRank Computation

- The Jacobi algorithm for solving linear systems: much faster than Gaussian elimination.

$$\boldsymbol{x}(t) = d\boldsymbol{W}\boldsymbol{x}(t-1) + (1-d)\mathbb{1}_N$$

How to calculate PR1? ⟶ PR1 = Transition Matrix * PR0

An iterative method

stable

To\From	WA	WB	WC	WD
WA	0	1/2	1	0
WB	1/3	0	0	1/2
WC	1/3	0	0	1/2
WD	1/3	1/2	0	0

	PR0
A	1/4
B	1/4
C	1/4
D	1/4

⟶

	PR1
A	9/24
B	5/24
C	5/24
D	5/24

PageRank Computation

- PageRank can be computed on graphs changing over time.

Normally, PageRank computation is based on a given fixed graph.

Provide a more general interpretation of PageRank in the case in which the graph changes over time.

Calculate the page scoring dynamically while the crawler is visiting the Web.

The sequence {**x**(t)} remains bounded, even if the transition matrix **W** = **W**(t) is updated during the computation of **x**(t).

PageRank Computation

- PageRank can be computed by an optimal algorithm

Billions of pages nowadays,

New challenge of tractability.

Let |H| be the number of hyperlinks in the Web. We prove that for a given degree of precision $\in$, the solution can be found with O(|H| log(1/ $\in$)) floating-point operations.

Conclusions

- Analysis of PageRank to disclose its fundamental properties: stability, complexity of computational scheme, and the critical role of parameters involved in the computation.
- Inherent structure of the Markovian matrices associated with the Web makes it possible to perform optimal computation of PageRank, a property of crucial importance for the actual scaling up the Web.
- The effectiveness of the computational scheme also depends on the limited precision requirements imposed by the sensitivity of PageRank from the dumping factor.
- Some nice properties concerning PageRank robustness have been derived.
- Technical soundness of a dynamical computational scheme.

Questions?