Efficient and Decentralized PageRank Approximation in a P2P Network

Josiane Xavier Parreira [⋆], Debora Donato [⋄], Sebastian Michel [⋆], Gerhard Weikum [⋆]

[⋆] Max-Planck Institute for Computer Science
[⋄] Università di Roma "La Sapienza"

September 13, 2006

Outline

Introduction

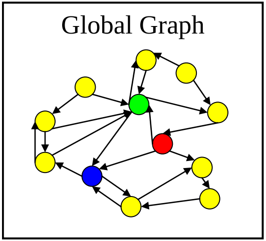
Computational Model

Every peer crawls Web fragments at its discretion and has its own local & personalized search engine

Introduction

Computational Model

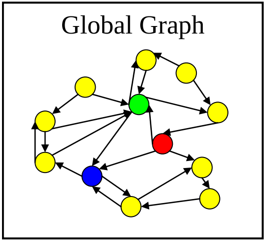
Every peer crawls Web fragments at its discretion and has its own local & personalized search engine



Introduction

Computational Model

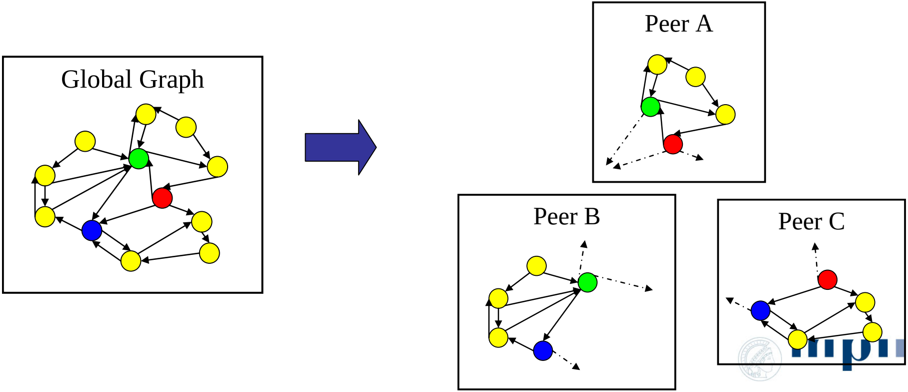
Every peer crawls Web fragments at its discretion and has its own local & personalized search engine



Introduction

Computational Model

Every peer crawls Web fragments at its discretion and has its own local & personalized search engine



Introduction

Goal

Compute "global" authority scores of pages in the network.

Introduction

Goal

Compute "global" authority scores of pages in the network.

Problems

- Peers have only local (incomplete) information
- Pages might link to or be linked by pages at other peers
- No control over overlaps between local graphs

PageRank

PageRank [Brin and Page, WWW'98]

- Importance of a page depends on the importance of the pages that point to it
- Stationary distribution of a Markov chain that describes a random walk over the graph
- Can be computed using the power iteration method

PageRank Formulation

$$PR(q) = \epsilon \times \sum_{p|p \rightarrow q} PR(p)/out(p) + (1-\epsilon) \times 1/N$$

Related Work

Efficient PR

- Graph Aggregation [Broder et al., WWW'04]
- Iterative Aggregation [Langville & Meyer, WWW'04]

Decentralized PR

- *Local PageRank & ServerRank* [Wang & DeWitt, VLDB'04]
- *BlockRank* [Kamvar et al., Stanford Tech. Report'03]

Markov Chains Aggregation/Disaggregation Techniques

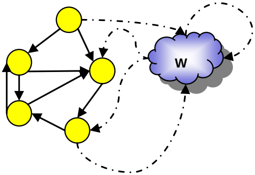
- Kemeny & Snell [1963]
- Stewart [1994]
- Meyer [2000]

JXP Algorithm

JXP Algorithm

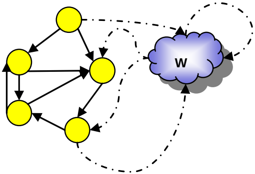
- Decentralized algorithm for computing authority scores of pages in a P2P Network, with arbitrary overlapping
- Runs locally at every peer
- No coordinator, asynchronous
- Combines local PageRank computations + Meetings between peers
- JXP scores converge to the true global PageRank scores

World Node



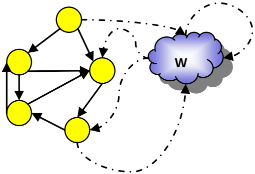
World Node

- Special node added to each local graph

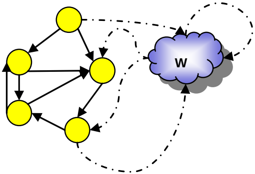


World Node

- Special node added to each local graph
- Represents all pages in the network that do not belong to local graph

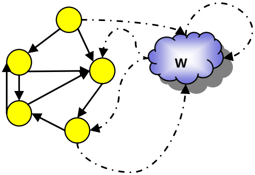


World Node



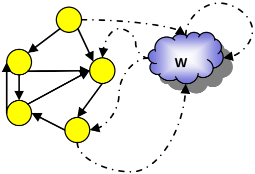
- Special node added to each local graph
- Represents all pages in the network that do not belong to local graph
- "Special features":

World Node



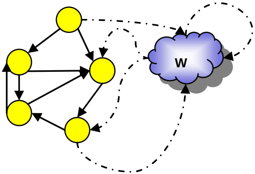
- Special node added to each local graph
- Represents all pages in the network that do not belong to local graph
- "Special features":
  - All links from local pages to external pages point to World Node

World Node



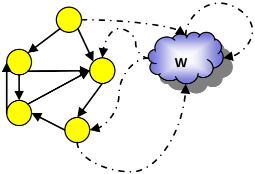
- Special node added to each local graph
- Represents all pages in the network that do not belong to local graph
- "Special features":
  - All links from local pages to external pages point to World Node
  - Links from external pages that point to local pages (discovered during meetings) are represented at the World Node

World Node



- Special node added to each local graph
- Represents all pages in the network that do not belong to local graph
- "Special features":
  - All links from local pages to external pages point to World Node
  - Links from external pages that point to local pages (discovered during meetings) are represented at the World Node
  - Score and outdegree of these external pages are stored; World Node outgoing links are weighted to reflect score mass given by original link

World Node



- Special node added to each local graph
- Represents all pages in the network that do not belong to local graph
- "Special features":
  - All links from local pages to external pages point to World Node
  - Links from external pages that point to local pages (discovered during meetings) are represented at the World Node
  - Score and outdegree of these external pages are stored; World Node outgoing links are weighted to reflect score mass given by original link
  - Self-loop link to represent transitions among external pages

The Algorithm

Initialization step

- Local graph is extended by adding the world node;
- PageRank is computed in the extended graph $\longrightarrow$ JXP scores

The Algorithm

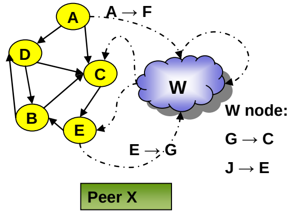
Initialization step

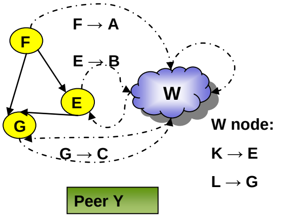
- Local graph is extended by adding the world node;
- PageRank is computed in the extended graph $\rightarrow$ JXP scores

Main Algorithm (for every P_i in the network)

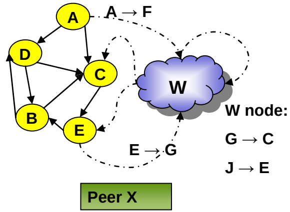
- Select P_j to meet
- Update world node
  - Add edges for pages in P_j that point to pages in P_i
  - If an edge already exists at the world node, the score of the source page is updated by taking the highest of both scores
- Compute PageRank $\rightarrow$ JXP scores

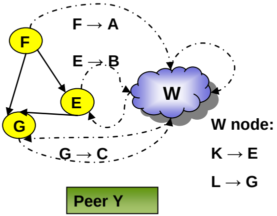
Example



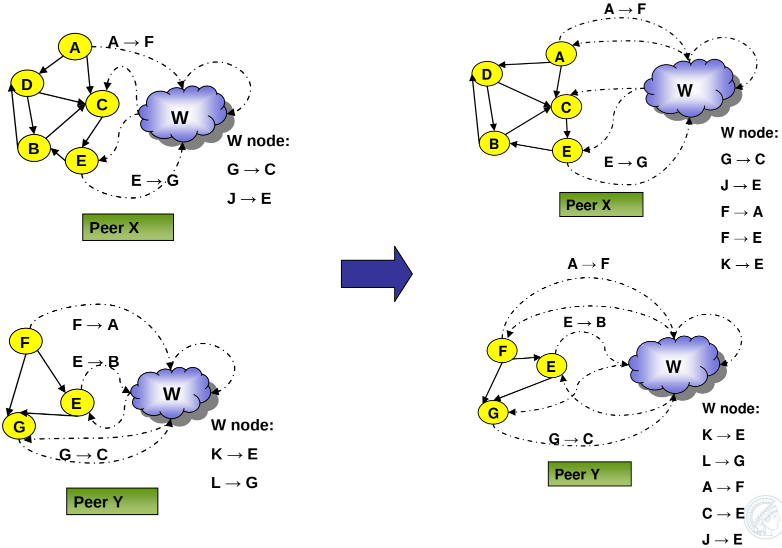


Example





Example


A → F
A
D
C
B
E
W
W node:
G → C
J → E
E → G
Peer X
F → A
F
E → B
E
G
W
W node:
K → E
L → G
G → C
Peer Y
A → F
A
D
C
B
E
W
W node:
G → C
J → E
F → A
F → E
K → E
E → G
Peer X
A → F
E → B
F
E
G
W
W node:
K → E
L → G
A → F
C → E
J → E
G → C
Peer Y

Peer Selection Strategy

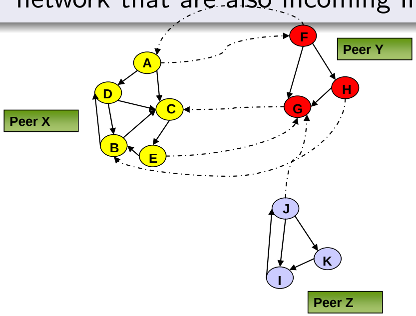
Motivation

- Peers' contribution for the convergence are different
- Finding peers with high contribution would speed up convergence
- "Quality indicator": Number of outgoing links of a peer in the network that are also incoming links in the local graph

Peer Selection Strategy

Motivation

- Peers' contribution for the convergence are different
- Finding peers with high contribution would speed up convergence
- "Quality indicator": Number of outgoing links of a peer in the network that are also incoming links in the local graph



Peer Selection Strategy

Good strategy

Find promising peers without increasing much bandwidth consumption

- Caching + statistical synopses

Peer Selection Strategy

Good strategy

Find promising peers without increasing much bandwidth consumption

- Caching + statistical synopses

Statistical synopses

Approximation technique for comparing data of different peers without explicitly transferring their contents.

- Compact representation of sets
- Can be used to estimate cardinality of the intersection between two sets
- JXP uses Min-Wise Independent Permutations (MIPs) [Broder et al., 1997]

Pre-meetings Strategy

- Each peer P_i computes *local*(P_i) and *successors*(P_i) MIPs vectors (256-integer vectors)
- When P_i meets P_j
  - Uses MIPs vectors to estimate percentage of local pages pointed by pages in P_j
  - If percentage above threshold, P_i caches P_j's ID
  - Uses MIPs again to estimate overlap between the two local graphs
  - If there is high overlap, peers exchange their list of cached ID's and store them in a temporary list
  - Idea: Peers on the temporary list are potential candidates for the next meeting

Pre-meetings Strategy

Pre-meetings phase

- P_j contacts peers on the temporary list and ask for their MIPs vectors
- Assign scores to each peer
- For next (real) meeting, P_i chooses P_k where
  - P_k is best scored peer in temporary list, with prob. α
  - P_k is one of the already cached peers, with prob. β
  - P_k is a random peer, with prob. $(1 - \alpha - \beta)$

Mathematical Analysis

Assumptions

Global transition matrix $\mathbf{C}_{N \times N}$ and global stationary distribution vector $\boldsymbol{\pi}$

Local transition matrix and local stationary distr. (JXP scores)

$$\mathbf{P} = \left(\begin{array}{ccc|c} p_{11} & \dots & p_{1n} & p_{1w} \\ \vdots & \dots & \vdots & \vdots \\ p_{n1} & \dots & p_{nn} & p_{nw} \\ \hline p_{w1} & \dots & p_{wn} & p_{ww} \end{array} \right)$$

$$p_{ij} = \begin{cases} \frac{1}{out(i)} & \text{if } \exists\, i \rightarrow j \\ 0 & \text{otherwise} \end{cases}$$

$$p_{iw} = \sum_{\substack{i \rightarrow r \\ r \notin G}} \frac{1}{out(i)}$$

$$\boldsymbol{\alpha} = \left(\begin{array}{ccc|c} \alpha_1 & \dots & \alpha_n & \alpha_w \end{array} \right)^T$$

for every i, j, $1 \leq i, j \leq n$. (G is the set of local pages)

Mathematical Analysis

World Node transitions prob.

$$p^t_{wi} = \sum_{\substack{r \rightarrow i \\ r \in W^t}} \frac{\alpha(r)^t}{out(r)} \cdot \frac{1}{\alpha^{t-1}_w} \qquad p^t_{ww} = 1 - \sum_{i=1}^{n} p^t_{wi}$$

W^t: Set of pages represented at the World Node during meeting t

Random Jumps

$$\mathbf{P}' = \epsilon\ \mathbf{P} + (1-\epsilon)\frac{1}{N}\left(\ 1 \quad \ldots \quad 1 \,\middle|\, (N-n)\ \right)$$

Mathematical Analysis

Meeting Step

Considering one link addition/update at a time

$$\mathbf{P}^t = \mathbf{P}^{t-1} + \mathbf{E} \quad \mathbf{E} = \left(\begin{array}{ccccccc|c} 0 & & & \dots & & & 0 & 0 \\ \vdots & & & \dots & & & \vdots & \vdots \\ 0 & & & \dots & & & 0 & 0 \\ \hline 0 & \dots & 0 & \delta & 0 & \dots & 0 & -\delta \end{array}\right)$$

Mathematical Analysis

Meeting Step

Considering one link addition/update at a time

$$\mathbf{P}^t = \mathbf{P}^{t-1} + \mathbf{E} \quad \mathbf{E} = \left(\begin{array}{ccccccc|c} 0 & & & \dots & & & 0 & 0 \\ \vdots & & & \dots & & & \vdots & \vdots \\ 0 & & & \dots & & & 0 & 0 \\ \hline 0 & \dots & 0 & \delta & 0 & \dots & 0 & -\delta \end{array}\right)$$

Theorem 1

The JXP score of the world node, at every peer in the network, is monotonically non-increasing.

Proof: Based on the study of the sensitivity of Markov Chains [Cho & Meyer, 1999].

Mathematical Analysis

Theorem 2

The sum of scores over all pages in a local graph, at every peer in the network, is monotonically non-decreasing.

Mathematical Analysis

Theorem 2

The sum of scores over all pages in a local graph, at every peer in the network, is monotonically non-decreasing.

Theorem 3

Consider the true stationary probabilities (PR scores) of pages $i \in G$ and the World Node w, π_i and π_w, and their JXP scores after t meetings α_i^t and α_w^t. The following holds throughout all JXP meetings:

$0 < \alpha_i^t \leq \pi_i$ *for* $i \in G$ *and* $\pi_w \leq \alpha_w^t < 1$.

Mathematical Analysis

Theorem 2

The sum of scores over all pages in a local graph, at every peer in the network, is monotonically non-decreasing.

Theorem 3

Consider the true stationary probabilities (PR scores) of pages $i \in G$ and the World Node w, π_i and π_w, and their JXP scores after t meetings α_i^t and α_w^t. The following holds throughout all JXP meetings:
$0 < \alpha_i^t \leq \pi_i$ *for* $i \in G$ *and* $\pi_w \leq \alpha_w^t < 1$.

Theorem 4

In a fair series of JXP meetings, the JXP scores of all nodes converge to the true global PR scores.

Setup

Amazon collection

- 55,196 pages
- 237,160 links
- 10 categories (e.g. Computers, Sports, Travel, etc)

Web collection

- 103,591 pages
- 1,633,276 links
- 10 categories (e.g. Movies, Music, Politics, etc)

Setup

Amazon collection

- 55,196 pages
- 237,160 links
- 10 categories (e.g. Computers, Sports, Travel, etc)

Web collection

- 103,591 pages
- 1,633,276 links
- 10 categories (e.g. Movies, Music, Politics, etc)

Setup

- 100 peers (10 peers/category)

Setup

Amazon collection

- 55,196 pages
- 237,160 links
- 10 categories (e.g. Computers, Sports, Travel, etc)

Web collection

- 103,591 pages
- 1,633,276 links
- 10 categories (e.g. Movies, Music, Politics, etc)

Setup

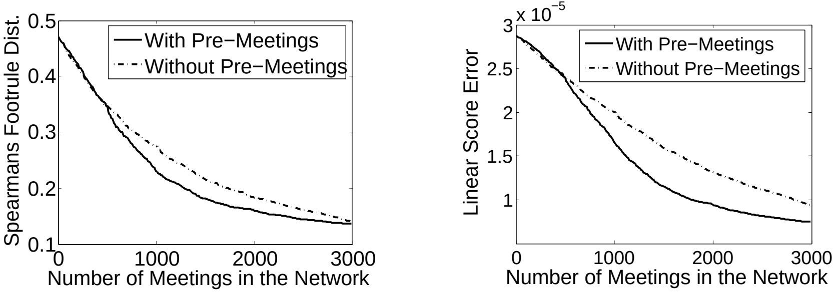
- 100 peers (10 peers/category)

Evaluation Measures

- "Global" JXP ranking vs. Global PageRank ranking
- Spearman's Footrule Distance at top-k
- Linear Score Error at top-k

Experimental Results

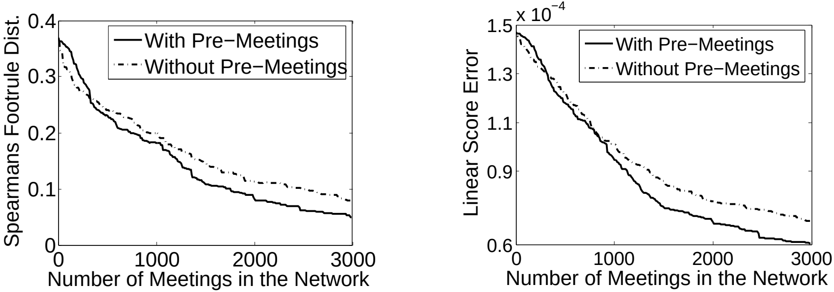
Amazon Collection, top-10000



For a footrule distance of 0.2 number of meetings was reduced from 1,770 to 1,250

Experimental Results

Web Collection, top-1000



For a footrule distance of 0.1 number of meetings was reduced from 2,480 to 1,650

Bandwidth Consumption

Web Collection

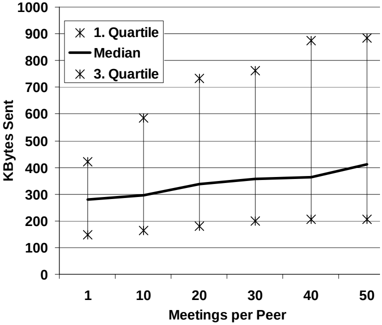


Figure: Without pre-meetings

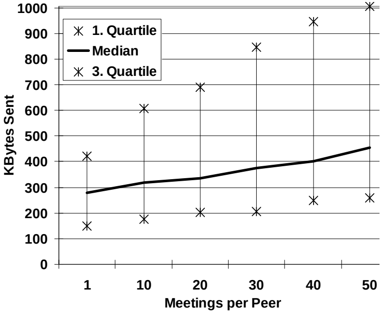


Figure: With pre-meetings

Message size (in KBytes) for the Web crawl setup

JXP in P2P Search

JXP integrated into our P2P search engine Minerva.
(Minerva Project Website: http://www.minerva-project.org)

Setup

- Bigger subset of Web (250,760 docs & 3,123,993 links)
- 40 peers, high overlap
- 15 queries [a], using the Minerva query routing mechanism
- Results were ranked in two ways:
  - tf*idf only
  - weighted sum of tf*idf and JXP scores
- Precision at top-10 measured (based on manually assessments)

[a] taken from Borodin et al., ACM TOIT, 2005

Results

Query	tf*idf	(0.6 tf*idf + 0.4 JXP)
affirmative action	40%	40%
amusement parks	60%	60%
armstrong	20%	**80%**
basketball	20%	**60%**
blues	20%	20%
censorship	**30%**	20%
cheese	40%	**60%**
iraq war	**50%**	30%
jordan	40%	40%
moon landing	**90%**	70%
movies	30%	**100%**
roswell	30%	**70%**
search engines	20%	**60%**
shakespeare	60%	**80%**
table tennis	50%	**70%**
Average	40%	**57%**

Conclusions and Ongoing Work

Conclusions

- JXP algorithm for dynamically computing authority scores of pages distributed in a P2P network
- Fully decentralized (no coordinator), asynchronous
- Combines local PageRank computation with meetings between peers
- JXP scores are proved to converge to global PageRank scores

Ongoing Work

- Integrate JXP into the query routing mechanism [P2PIR'06]
- JXP in dynamic networks
- Adapt JXP to work in the presence of malicious peers