Robust PageRank and Locally Computable Spam Detection Features

Vahab Mirrokni [Microsoft Research]
–joint work with–
Reid Andersen [Microsoft Research]
Christian Borgs [Microsoft Research]
Jennifer Chayes [Microsoft Research]
John Hopcroft [Cornell University]
Kamal Jain [Microsoft Research]
Shang-Hua Teng [Boston University]

Outline

- PageRank and PageRank Contributions.
- Applications to Link Spam Detection.
- A Local Algorithm for PageRank Contributions.
- Link Spam Detection Features and Experimental Results.

PageRank

PageRank measures the importance of nodes in a graph.

PageRank on the web graph:

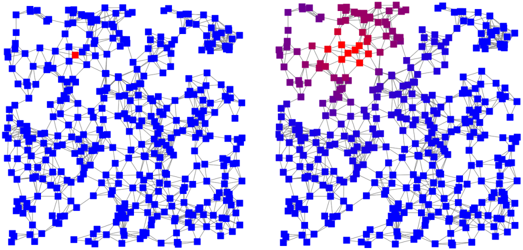
- Rank pages for a query.
- Priority in web crawls.

PageRank:

- Link Structure.
- PageRank score depends recursively on the PageRank score of incomming neighbors.

Where does the PageRank come from?

PageRank score: the sum of the *PageRank contributions* from other nodes.



Outgoing contributions: Each node sends small contributions to the nodes it can reach either directly or indirectly.
Incoming contributions: The PageRank of a particular node is the sum of the contributions it receives.

Definition of PageRank with an arbitrary starting distribution

We now define the PageRank vector $\mathbf{pr}(\alpha, s)$.

s is an arbitrary *restarting distribution*
α is the *restarting probability*.

Definition of PageRank

Consider the following random walk in the graph. At each step:

$$\begin{cases} \text{move to a neighbor at random with probablity } (1-\alpha) \\ \text{restart to } s \text{ with probablity } \alpha. \end{cases}$$

PageRank $\mathbf{pr}(\alpha, s)[v]$ is the stable distribution of the above random walk.

Global PageRank and Personalized PageRank

These are special cases of PageRank, with specific starting distributions.

Personalized PageRank

In personalized PageRank for u, $s = \mathbf{e}_u$ (vector with a one at u).

Global PageRank (the usual PageRank)

In PageRank, $s = \mathbf{1}$.

Relationship between the two

Global PageRank vector = the sum of the personalized PageRank vectors.

Global PageRank and Personalized PageRank

These are special cases of PageRank, with specific starting distributions.

Personalized PageRank

In personalized PageRank for u, $s = \mathbf{e}_u$ (vector with a one at u).

Global PageRank (the usual PageRank)

In PageRank, $s = \mathbf{1}$.

Relationship between the two

Global PageRank vector = the sum of the personalized PageRank vectors.

Definition

The *contribution* from u to v = the personalized PageRank of u for v.

Link Spam and PageRank Contribution

Link Spam: Web Spammers abuse the link structure and get high PageRank without introducing new content.

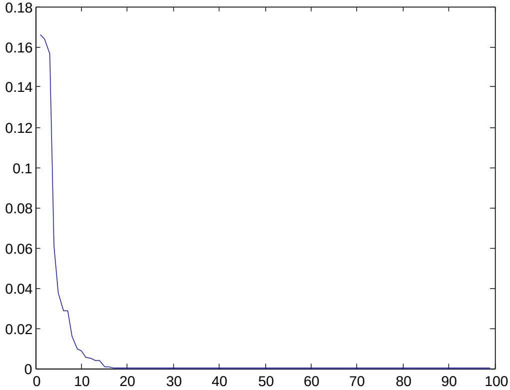
Link Spam and PageRank Contribution

Link Spam: Web Spammers abuse the link structure and get high PageRank without introducing new content.

- The PageRank of a high PageRank non-spam node consists of small contributions from a large set of nodes.
- The PageRank of a high PageRank spam node consists of large contributions from a small set of nodes.
- This has been formally observed by SpamRank [Benczur et al. 05].

Plot of contributions

Contribution Vector of a spam node from UK host graph.

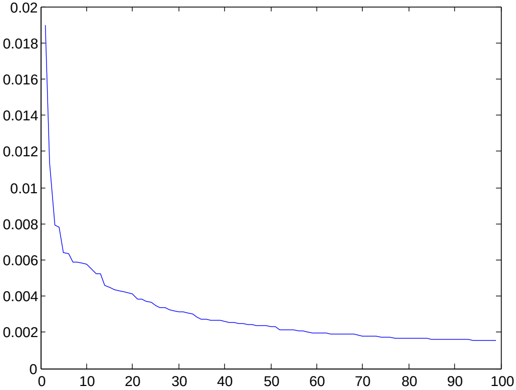


X-axis: Node Number (sorted by contribution)

Y-axis: Contribution

Plot of contributions

Contribution Vector of a non-spam node from UK host graph.



X-axis: Node Number (sorted by contribution)

Y-axis: Contribution

Identifying top contributors

Problem: Given a page, identify its top contributors.

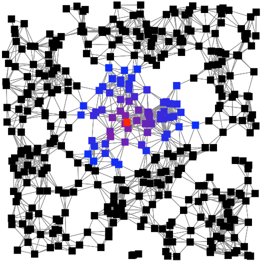
- Identify the top k contributors.
- Identify all pages who contribute above a certain threshold (i.e, they have large personalized PageRank to this page).

Our Goal: Approximate the contribution vector to a node, using a local algorithm.

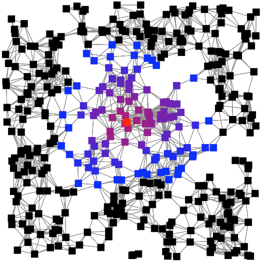
Local Algorithm:

- It examines only a small part of the entire graph.
- It produces a sparse approximate solution. It produces an *approximate contribution vector* that differs from the true vector of contributions by at most ϵ at each node.

Approximate contributions for different ϵ.



$\epsilon = .001$



$\epsilon = .0005$

Applications for locally computing PageRank contributions

Locally Computable Link Spam Features.
Supervised and Unsupervised features can be computed on-the-fly for a few selected nodes.

Support of Known Spam Pages.
For a spam node, identify its top contributors.

Description of the contribution algorithm

Technique: The *asynchronous pushing method* [Jeh/Widom 03] [McSherry 05] [Berkhin 06]

Description of the contribution algorithm

Technique: The *asynchronous pushing method* [Jeh/Widom 03] [McSherry 05] [Berkhin 06]

The algorithm `ApproxContributions`(v, α, ϵ)

Input:
v, the target node
α, the PageRank restarting probability
ϵ, the desired error in each entry of the contribution vector
Output: An ϵ-approximate contribution vector

Description of the contribution algorithm

Technique: The *asynchronous pushing method* [Jeh/Widom 03] [McSherry 05] [Berkhin 06]

The algorithm `ApproxContributions`(v, α, ϵ)

Input:
v, the target node
α, the PageRank restarting probability
ϵ, the desired error in each entry of the contribution vector
Output: An ϵ-approximate contribution vector

`pushback` Operation: Push some probability to each in-neighbor.

Description of the contribution algorithm

Technique: The *asynchronous pushing method* [Jeh/Widom 03] [McSherry 05] [Berkhin 06]

The algorithm `ApproxContributions`(v, α, ϵ)

Input:
v, the target node
α, the PageRank restarting probability
ϵ, the desired error in each entry of the contribution vector
Output: An ϵ-approximate contribution vector

`pushback` Operation: Push some probability to each in-neighbor.

Running Time: The number of pushback operations is linear in size of contribution set.

Computing contributions locally

We will maintain two vectors,

- an ϵ-approximate contribution vector $\mathbf{c}$, and
- a residual vector $\mathbf{r}$.

Computing contributions locally

We will maintain two vectors,

- an ϵ-approximate contribution vector $\mathbf{c}$, and
- a residual vector $\mathbf{r}$.

pushback(c,r,u)

```
c'[u] += alpha * r[u]
r'[u] = 0
for v such that v -> u:
    r'[v] += (1-alpha) r[u] / outdegree[v]
change r to r' and c to c'
```

Computing contributions locally

We will maintain two vectors,

- an ϵ-approximate contribution vector **c**, and
- a residual vector **r**.

pushback(c,r,u)

```
c'[u] += alpha * r[u]
r'[u] = 0
for v such that v -> u:
    r'[v] += (1-alpha) r[u] / outdegree[v]
change r to r' and c to c'
```

Main Loop

While there is a node u where $r(u) > \epsilon$,
pick any such node and perform the push operation.

Example: identifying sets of top contributors locally

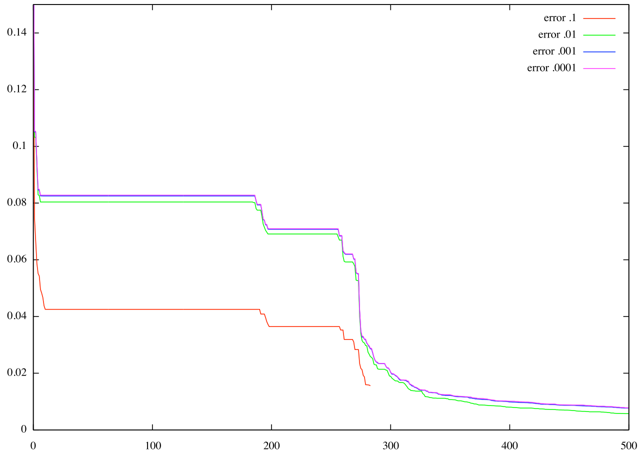
target page: www.usajobs.opm.gov/b.htm
desired error: $\epsilon = .001$

Top contributors and their approximate contributions

0.206109	www.usajobs.opm.gov/b.htm
0.105105	www.rurdev.usda.gov/rbs/oa/jobs.htm
0.105105	www.fsa.usda.gov/pas/fsajobs.htm
0.0946422	staffing.opm.gov/Immigrationinspector/
0.0846548	www.usajobs.opm.gov/survey.htm
0.0845882	profiler.usajobs.opm.gov/
0.0825384	www.usajobs.opm.gov/a9nasa.htm
0.0825384	www.usajobs.opm.gov/a9noaa.htm
0.0825086	www.usajobs.opm.gov/wfjic/jobs/T04034.htm
0.0825086	www.usajobs.opm.gov/wfjic/jobs/IA2386.htm
0.0825086	www.usajobs.opm.gov/wfjic/jobs/IZ9687.htm
0.0825086	www.usajobs.opm.gov/wfjic/jobs/IZ9590.htm

- The algorithm examines 5877 vertices.
- It finds 1777 pages that contribute at least ϵ to the target.
- It produces an approximate contribution vector where the error in each entry is at most $\epsilon = 0.001$.

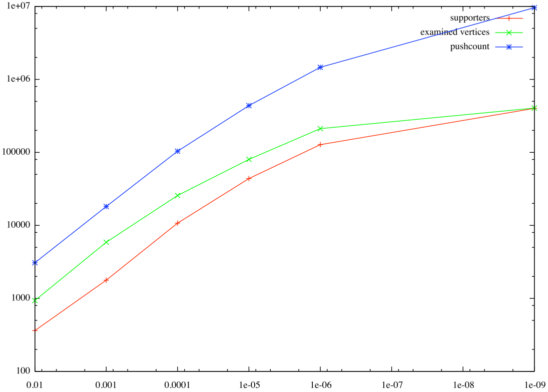
Approximate contributions



X-axis: Node Number (sorted by contribution)

Y-axis: Contribution (lower bounds with some error)

Running Time



log-log plot
X-axis: Error level ϵ in the contribution vector
Y-axis: number of ϵ-supporters and number of nodes examined

Locally Computable Link Spam Features

Definition

$S_\delta(v)$: The δ-contributing set of a node v is the set of nodes whose contributions to v are at least δpr(v).

Locally Computable Link Spam Features

Definition

$S_\delta(v)$: The δ-contributing set of a node v is the set of nodes whose contributions to v are at least δpr(v).

Supervised features:

- Ratio of spam in contributing set: Ratio of spam and Non-spam nodes in the δ-contributing set.

Locally Computable Link Spam Features

Definition

$S_\delta(v)$: The δ-contributing set of a node v is the set of nodes whose contributions to v are at least δpr(v).

Supervised features:

- Ratio of spam in contributing set: Ratio of spam and Non-spam nodes in the δ-contributing set.

Unsupervised features:

- Size of the δ-contributing set ($|S_\delta(v)|$).
- l_1 and l_2 norm of contribution vector of the δ-contributing set.

Robust PageRank

Robust PageRank = sum of truncated contributions.

Generalize Truncated PageRank by Becchetti et al. (2006).

Robust PageRank

Robust PageRank = sum of truncated contributions.
Generalize Truncated PageRank by Becchetti et al. (2006).

$$\begin{aligned}
\mathrm{Robustpr}_\alpha^\delta(v) &= \sum_{u \in V(G)} \min(\mathbf{ppr}(u,v), \delta) \\
&= \sum_{u \in V(G)} \mathbf{ppr}(u,v) - \sum_{u \in S_\delta(v)} (\mathbf{ppr}(u,v) - \delta) \\
&= \mathrm{pr}_\alpha(v) - \sum_{u \in S_\delta(v)} \mathbf{ppr}(u,v) - \delta |S_\delta(v)|.
\end{aligned}$$

Robust PageRank

Robust PageRank = sum of truncated contributions.
Generalize Truncated PageRank by Becchetti et al. (2006).

$$\begin{aligned}\mathrm{Robustpr}_{\alpha}^{\delta}(v) &= \sum_{u \in V(G)} \min(\mathbf{ppr}(u,v), \delta) \\ &= \sum_{u \in V(G)} \mathbf{ppr}(u,v) - \sum_{u \in S_\delta(v)} (\mathbf{ppr}(u,v) - \delta) \\ &= \mathrm{pr}_{\alpha}(v) - \sum_{u \in S_\delta(v)} \mathbf{ppr}(u,v) - \delta |S_\delta(v)|.\end{aligned}$$

Feature: Ratio between Robust PageRank and PageRank.

Performance of Link Spam Features

Labeled UK Host Graph

11401 nodes, average degree 65,

Examined 24% high PageRank nodes

$\delta = 10^{-4}$, average size of δ-contributing set= 301

Feature	FNeg1	FPos1	FNeg2	FPos2

Performance of Link Spam Features

Labeled UK Host Graph

11401 nodes, average degree 65,

Examined 24% high PageRank nodes

$\delta = 10^{-4}$, average size of δ-contributing set$= 301$

Feature	FNeg1	FPos1	FNeg2	FPos2
Size	8%	5%	78%	2%
l_1 Norm	6%	5%	67%	2%
$\frac{\text{Robust PR}}{\text{PR}}$	5%	5%	38%	2%

Performance of Link Spam Features

Labeled UK Host Graph

11401 nodes, average degree 65,

Examined 24% high PageRank nodes

$\delta = 10^{-4}$, average size of δ-contributing set= 301

Feature	FNeg1	FPos1	FNeg2	FPos2
Size	8%	5%	78%	2%
l_1 Norm	6%	5%	67%	2%
$\frac{\text{Robust PR}}{\text{PR}}$	5%	5%	38%	2%
Indegree (Base)	45%	5%	78%	2%
PRIndegree (Base)	50%	5%	82%	2%

Performance of Link Spam Features

Labeled UK Host Graph

11401 nodes, average degree 65,

Examined 24% high PageRank nodes

$\delta = 10^{-4}$, average size of δ-contributing set= 301

Feature	FNeg1	FPos1	FNeg2	FPos2
Size	8%	5%	78%	2%
l_1 Norm	6%	5%	67%	2%
$\frac{\text{Robust PR}}{\text{PR}}$	5%	5%	38%	2%
Indegree (Base)	45%	5%	78%	2%
PRIndegree (Base)	50%	5%	82%	2%
Spam in Contrib. (Sup)	4%	5%	15%	2%
Spam in Neighbors (Base)	8%	5%	33%	2%

Other Related Work

- Topic-sensitive PageRank [Haveliwala 03],
- TrustRank [Gyongyi et al. 04],
- Anti-TrustRank [Raj et al. 99],
- SpamMass algorithm [Gyongyi et al. 06].

- Estimating PageRank.
  The PageRank of a node can be estimated within a smaller subgraph containing its large contributors [Chen et al. 04].

Thank You