Using rank propagation and Probabilistic counting for Link-Based Spam Detection

Luca Becchetti\textsuperscript{1}, Carlos Castillo\textsuperscript{1}, Debora Donato\textsuperscript{1}, Stefano Leonardi\textsuperscript{1} and Ricardo Baeza-Yates\textsuperscript{2}

1. Università di Roma “La Sapienza” – Rome, Italy
2. Yahoo! Research – Barcelona, Spain and Santiago, Chile

August 20th, 2006
Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation

1. Motivation
2. Spam pages characterization
3. Truncated PageRank
4. Counting supporters
5. Experiments
6. Conclusions
Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

What is on the Web?

Information + Porn
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

What is on the Web?

Information + Porn + Online casinos + Free movies + Cheap software + Buy a MBA diploma + Prescription-free drugs + V!-4-gra + Get rich now now now now!!!

Graphic: www.milliondollarhomepage.com
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Web spam (keywords + links)

Viagra for women, has viagra cost lowest prices viagra, at cialis eli lilly, non prescription viagra, am cialis on line, viagra for women viagra expiration cialis fda approval, compare viagra and levitra viagra discount viagra cialis levitra, viagra online cheap cialis no prescription, 180 mg viagra levitra vs viagra uk viagra viagra sample, am generic cialis minuteviagra cum-with-us.com, free viagra online.

Herbal viagra samples, to order viagra visit your doctor online viagra substitute side effects from viagra cheapest price viagra, by cialis soft tab, mail order viagra, for cialis store, british viagra, is cialis fedex overnight, viagra suppliers cialis herbalsubstitute com, whats the chemical name for the drug viagra, herbal viagra, viagra info.
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation
Spam pages characterization
Truncated PageRank
Counting supporters
Experiments
Conclusions

Web spam (mostly keywords)

smart movie converter 2.72 registration key as nokia6600
crac diablo2, kontakt 2 .exe, download crack norton 2006 online, dowload snagit, telecharger canopu
nothing else mather/lyric, silent hill 3 no-dvd crack, Protocol v7 Update-Vengeance.rar download, kn
windows validation crack download
ftp downloads spanish,
total commander hack,
plus superpack key,
crack de need ford
speed underground,
manual diablo 2,
Rapture_dhol_mix.mp3,
advance 3gp convertor,
pacific assault torrent,
Fruity Loops 4.5.1
Demo, russian mohaa
skin, telechargement de

key generator for easy cd-da extractor v9, fine rider
8.0, donwload demo fifa street pc, soundtreck
moulin rouge free mp3, crazy froog popcorn, Utah
Saints - Take On The Theme From Mortal Kombat
mp3, cunter-strike password, download free game
sex, SYSTEM OF DOWN DIRECTORY
PARENT MPEG

wifi download key generator wap, speedstream
feature activation, command conquer generals key
gens, nerovision directx9.0 download, swift 3D trial

free gemu,
telecharger
cdkeys, care day
home question,
Remote S60
software

spells, craked, nt print
server,
SWF2Video
Plugin for
Adobe Premiere
Pro craked, ps2
secret code
download
webcam
lv-c300,
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Fake search engine

Results 1-16 containing "1293kasd132ka0sd1kj239asd123"

1. A Real Work At Home Business Opportunity!
Free Home Business Match Up Service! We have helped 1000's of people make $5,000...
http://gozing.directtrack.com/z/1198/CD2127/

2. Exotic Holiday - Find Your Love
Exotic holiday is great way how to find love when you travel. Meet new people. Meet...
http://www.exotic-holiday.co.uk/

3. Image, Photo, Digital, Video and Movie software
Find quality image management &amp; digital asset software for your business. Also so...
http://www.enterprise-software.co.uk

4. Renting a Birthday Party Limousine is Sexy
What better way to surprise your loved one on their special day than with a birthday p...
http://partybusrental.info
Website design, management, marketing and promotion

If you are searching for any of the following topics:

- Website design, management, marketing and promotion
- Website design, management, marketing and promotion resources
- Website design, management, marketing and promotion related topics
- Website design, management, marketing and promotion services

Look No further. You'll find it at Website design, management, marketing and promotion

Website design, management, marketing and promotion is the key to your needs. You're one step ahead with Dry Media.

Website design, management, marketing and promotion brought to you by Dry Media, the leaders in this field.

At the Website design, management, marketing and promotion web site, you'll discover an easy to use, information packed source of data on Website design, management, marketing and promotion.

Click Here to Learn More about Website design, management, marketing and promotion.
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation

Problem: “normal” pages that are spam

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions
Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Problem: borderline pages
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

1 Motivation
2 Spam pages characterization
3 Truncated PageRank
4 Counting supporters
5 Experiments
6 Conclusions
Definitions

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Any deliberate action that is meant to trigger an unjustifiably favorable relevance or importance for some Web page, considering the page's true value [Gyöngyi and Garcia-Molina, 2005] any attempt to deceive a search engine's relevancy algorithm or simply anything that would not be done if search engines did not exist. [Perkins, 2001]
Definitions

Any deliberate action that is meant to trigger an unjustifiably favorable relevance or importance for some Web page, considering the page’s true value [Gyöngyi and Garcia-Molina, 2005]
Definitions

Any deliberate action that is meant to trigger an unjustifiably favorable relevance or importance for some Web page, considering the page’s true value [Gyöngyi and Garcia-Molina, 2005]

any attempt to deceive a search engine’s relevancy algorithm or simply

anything that would not be done if search engines did not exist. [Perkins, 2001]
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Link farms

Single-level farms can be detected by searching groups of nodes sharing their out-links [Gibson et al., 2005]
Single-level farms can be detected by searching groups of nodes sharing their out-links [Gibson et al., 2005]
[Fetterly et al., 2004] hypothesized that studying the distribution of statistics about pages could be a good way of detecting spam pages:

“in a number of these distributions, outlier values are associated with web spam”
[Fetterly et al., 2004] hypothesized that studying the distribution of statistics about pages could be a good way of detecting spam pages:

“in a number of these distributions, outlier values are associated with web spam”

Research goal

Statistical analysis of link-based spam
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation
Spam pages characterization
Truncated PageRank
Counting supporters
Experiments
Conclusions

Idea: count “supporters” at different distances

Number of different nodes at a given distance:

**.UK** 18 mill. nodes

![Graph showing frequency distribution for .UK]

Average distance 14.9 clicks

**.EU.INT** 860,000 nodes

![Graph showing frequency distribution for .EU.INT]

Average distance 10.0 clicks
High and low-ranked pages are different

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Areas below the curves are equal if we are in the same strongly-connected component
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation
Spam pages characterization
Truncated PageRank
Counting supporters
Experiments
Conclusions

High and low-ranked pages are different

Areas below the curves are equal if we are in the same strongly-connected component
Motivation
Spam pages characterization
Truncated PageRank
Counting supporters
Experiments
Conclusions

---

**Graph algorithms**

All shortest paths, centrality, betweenness, clustering coefficient...

**Streamed algorithms**

Breadth-first and depth-first search
Count of neighbors

**Symmetric algorithms**

(Strongly) connected components
Approximate count of neighbors
PageRank, Truncated PageRank, Linear Rank
HITS, Salsa, TrustRank
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation
Spam pages characterization
Truncated PageRank
Counting supporters
Experiments
Conclusions
General functional ranking

Let \( P \) the row-normalized version of the citation matrix of a graph \( G = (V, E) \)

A **functional ranking** [Baeza-Yates et al., 2006] is a link-based ranking algorithm to compute a scoring vector \( W \) of the form:

\[
W = \sum_{t=0}^{\infty} \frac{\text{damping}(t)}{N} P^t.
\]
General functional ranking

Let $P$ the row-normalized version of the citation matrix of a graph $G = (V, E)$

A **functional ranking** [Baeza-Yates et al., 2006] is a link-based ranking algorithm to compute a scoring vector $W$ of the form:

$$W = \sum_{t=0}^{\infty} \frac{\text{damping}(t)}{N} P^t.$$

There are many choices for $\text{damping}(t)$, including simply a linear function that is as good as PageRank in practice.
Let $P$ the row-normalized version of the citation matrix of a graph $G = (V, E)$.

A functional ranking [Baeza-Yates et al., 2006] is a link-based ranking algorithm to compute a scoring vector $W$ of the form:

$$W = \sum_{t=0}^{\infty} \frac{damping(t)}{N} P^t.$$

There are many choices for $damping(t)$, including simply a linear function that is as good as PageRank in practice:

$$damping(t) = (1 - \alpha) \alpha^t$$
Using rank propagation and Probabilistic counting for Link-Based Spam Detection
L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation
Spam pages characterization
Truncated PageRank
Counting supporters
Experiments
Conclusions

Truncated PageRank

Reduce the direct contribution of the first levels of links:

\[
damping(t) = \begin{cases} 
0 & t \leq T \\
C\alpha^t & t > T 
\end{cases}
\]
Reduce the direct contribution of the first levels of links:

$$damping(t) = \begin{cases} 
0 & t \leq T \\
C\alpha^t & t > T 
\end{cases}$$

✓ No extra reading of the graph after PageRank
General algorithm

**Require:** N: number of nodes, 0 < α < 1: damping factor, T ≥ −1: distance for truncation

1: **for** i : 1 . . . N do  {Initialization}
2: R[i] ← (1 − α)/(α^{T+1}N)
3: **if** T ≥ 0 then
4:    Score[i] ← 0
5: **else** {Calculate normal PageRank}
6:    Score[i] ← R[i]
7: **end if**
8: **end for**

9: for i : 1 . . . N do  {Initialization}
10: R[i] ← (1 − α)/(α^{T+1}N)
11: **if** T ≥ 0 then
12:    Score[i] ← 0
13: **else** {Calculate normal PageRank}
14:    Score[i] ← R[i]
15: **end if**
16: **end for**
17: distance = 1
18: while not converged do
19:    Aux ← 0
20:    **for** src : 1 . . . N do  {Follow links in the graph}
21:        **for all** link from src to dest do
23:        **end for**
24:    **end for**
25:    **for** i : 1 . . . N do  {Apply damping factor}
26:        R[i] ← Aux[i] × α
27:        **if** distance > T then
28:            Score[i] ← Score[i] + R[i]
29:        **end if**
30:    **end for**
31:    distance = distance + 1
32: **end while**
33: return Score
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

General algorithm

**Require:** $N$: number of nodes, $0 < \alpha < 1$: damping factor, $T \geq -1$: distance for truncation

1: for $i : 1 \ldots N$ do \{Initialization\}
2: \hspace{0.5cm} $R[i] \leftarrow (1 - \alpha)/(\alpha^{T+1}N)$
3: \hspace{0.5cm} if $T \geq 0$ then
4: \hspace{1.0cm} $Score[i] \leftarrow 0$
5: \hspace{0.5cm} else \{Calculate normal PageRank\}
6: \hspace{1.0cm} $Score[i] \leftarrow R[i]$
7: \hspace{0.5cm} end if
8: end for
9: distance = 1
10: while not converged do
11: \hspace{0.5cm} Aux $\leftarrow 0$
12: \hspace{0.5cm} for src : $1 \ldots N$ do \{Follow links in the graph\}
13: \hspace{1.0cm} for all link from src to dest do
14: \hspace{1.5cm} Aux[dest] $\leftarrow$ Aux[dest] + $R[src]/\text{outdegree}(src)$
15: \hspace{1.0cm} end for
16: \hspace{0.5cm} end for
17: \hspace{0.5cm} for $i : 1 \ldots N$ do \{Apply damping factor $\alpha$\}
18: \hspace{1.0cm} $R[i] \leftarrow Aux[i] \times \alpha$
19: \hspace{1.0cm} if distance > $T$ then \{Add to ranking value\}
20: \hspace{1.5cm} $Score[i] \leftarrow Score[i] + R[i]$
21: \hspace{1.0cm} end if
22: \hspace{0.5cm} end for
23: \hspace{0.5cm} distance = distance + 1
24: end while
Comparing PageRank and Truncated PageRank with $T = 1$ and $T = 4$.
The correlation is high and decreases as more levels are truncated.
Motivation
Spam pages characterization
Truncated PageRank
Counting supporters
Experiments
Conclusions
Probabilistic counting

Motivation
Spam pages characterization
Counting supporters
Experiments
Conclusions
Improvement of ANF algorithm [Palmer et al., 2002] based on probabilistic counting [Flajolet and Martin, 1985]
**General algorithm**

**Require:** \( N \): number of nodes, \( d \): distance, \( k \): bits

1. **for** node : 1 \( \ldots \) N, bit: 1 \( \ldots \) k **do**
2. \texttt{INIT}(node,bit)
3. **end for**
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation
Spam pages characterization
Counting supporters
Experiments
Conclusions

General algorithm

Require: N: number of nodes, d: distance, k: bits

1: for node : 1 . . . N, bit: 1 . . . k do
2: INIT(node,bit)
3: end for
4: for distance : 1 . . . d do \{Iteration step\}
5: Aux ← 0_k
6: for src : 1 . . . N do \{Follow links in the graph\}
7: for all links from src to dest do
8: Aux[dest] ← Aux[dest] OR V[src,·]
9: end for
10: end for
11: V ← Aux
12: end for
General algorithm

**Require:** N: number of nodes, d: distance, k: bits
1. for node : 1 ... N, bit: 1 ... k do
2. INIT(node,bit)
3. end for
4. for distance : 1 ... d do {Iteration step}
5. Aux ← 0<sub>k</sub>
6. for src : 1 ... N do {Follow links in the graph}
7. for all links from src to dest do
9. end for
10. end for
11. V ← Aux
12. end for
13. for node: 1 ... N do {Estimate supporters}
14. Supporters[node] ← ESTIMATE( V[node,·] )
15. end for
16. return Supporters
Our estimator

Initialize all bits to one with probability $\epsilon$
Our estimator

Initialize all bits to one with probability $\epsilon$ by the independence of the $i$–th component $X_i$‘s we have,

$$P[X_i = 1] = 1 - (1 - \epsilon)^{\text{neighbors}(\text{node})},$$
Our estimator

Initialize all bits to one with probability $\epsilon$
by the independence of the $i-th$ component $X_i$’s we have,

$$P[X_i = 1] = 1 - (1 - \epsilon)^{\text{neighbors}(node)},$$

Estimator: $\text{neighbors}(node) = \log_{(1-\epsilon)}\left(1 - \frac{\text{ones}(node)}{k}\right)$
Our estimator

Initialize all bits to one with probability $\epsilon$ by the independence of the $i-th$ component $X_i$’s we have,

$$P[X_i = 1] = 1 - (1 - \epsilon)^{\text{neighbors}(node)},$$

Estimator: $\text{neighbors}(node) = \log_{(1-\epsilon)} \left( 1 - \frac{\text{ones}(node)}{k} \right)$

Problem: $\text{neighbors}(node)$ can vary by orders of magnitudes as $node$ varies.
Our estimator

Initialize all bits to one with probability $\epsilon$
by the independence of the $i$–th component $X_i$'s we have,

$$P[X_i = 1] = 1 - (1 - \epsilon)^{\text{neighbors}(\text{node})},$$

Estimator: $\text{neighbors}(\text{node}) = \log_{1-\epsilon} \left(1 - \frac{\text{ones}(\text{node})}{k}\right)$

**Problem:** $\text{neighbors}(\text{node})$ can vary by orders of magnitudes as $\text{node}$ varies.

This means that for some values of $\epsilon$, the computed value of $\text{ones}(\text{node})$ might be $k$ (or 0, depending on $\text{neighbors}(\text{node})$) with relatively high probability.
if we knew neighbors\((node)\) and chose \(\epsilon = \frac{1}{\text{neighbors}(node)}\) we would get:

\[
\text{ones}(node) \approx \left(1 - \frac{1}{e}\right)^k \approx 0.63k,
\]

Adaptive estimation

Repeat the above process for \(\epsilon = 1/2, 1/4, 1/8, \ldots\), and look for the transitions from more than \((1 - 1/e)k\) ones to less than \((1 - 1/e)k\) ones.
Using rank propagation and probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation
Spam pages characterization
Truncated PageRank
Counting supporters
Experiments
Conclusions

Convergence

![Graph showing convergence of fraction of nodes with estimates over iterations for different distances.](image)

- **Convergence**: 15 iterations for estimating the neighbors at distance 4 or less, less than 25 iterations for all distances up to 8.
15 iterations for estimating the neighbors at distance 4 or less
15 iterations for estimating the neighbors at distance 4 or less less than 25 iterations for all distances up to 8.
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

Error rate

![Graph showing average relative error vs distance]

- **Ours 64 bits, epsilon-only estimator**
- **Ours 64 bits, combined estimator**
- **ANF 24 bits × 24 iterations (576 b×i)**
- **ANF 24 bits × 48 iterations (1152 b×i)**
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation
Spam pages characterization
Truncated PageRank
Counting supporters
Experiments
Conclusions
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation
Spam pages characterization
Truncated PageRank
Counting supporters
Experiments
Conclusions

Test collection

**U.K. collection**

18.5 million pages downloaded from the .UK domain

5,344 hosts manually classified (6% of the hosts)
### Test collection

<table>
<thead>
<tr>
<th>U.K. collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.5 million pages downloaded from the .UK domain</td>
</tr>
<tr>
<td>5,344 hosts manually classified (6% of the hosts)</td>
</tr>
</tbody>
</table>

**Classified entire hosts:**

- A few hosts are mixed: spam and non-spam pages
- More coverage: sample covers 32% of the pages
Automatic classifier

We extracted (for the home page and the page with maximum PageRank) PageRank, Truncated PageRank at 2...4, Supporters at 2...4.

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions
Automatic classifier

We extracted (for the home page and the page with maximum PageRank) PageRank, Truncated PageRank at 2 . . . 4, Supporters at 2 . . . 4

We measured:

\[
\text{Precision} = \frac{\text{# of spam hosts classified as spam}}{\text{# of hosts classified as spam}}
\]

\[
\text{Recall} = \frac{\text{# of spam hosts classified as spam}}{\text{# of spam hosts}}.
\]
We extracted (for the home page and the page with maximum PageRank) PageRank, Truncated PageRank at 2...4, Supporters at 2...4.

We measured:

\[
\text{Precision} = \frac{\text{\# of spam hosts classified as spam}}{\text{\# of hosts classified as spam}}
\]

\[
\text{Recall} = \frac{\text{\# of spam hosts classified as spam}}{\text{\# of spam hosts}}.
\]

and the two types of errors in spam classification:

\[
\text{False positive rate} = \frac{\text{\# of normal hosts classified as spam}}{\text{\# of normal hosts}}
\]

\[
\text{False negative rate} = \frac{\text{\# of spam hosts classified as normal}}{\text{\# of spam hosts}}.
\]
Classifier based on TrustRank: uses as features the PageRank, the estimated non-spam mass, and the estimated non-spam mass divided by PageRank.

Classifier based on Truncated PageRank: uses as features the PageRank, the Truncated PageRank with truncation distance $t = 2, 3, 4$ (with $t = 1$ it would be just based on in-degree), and the Truncated PageRank divided by PageRank.

Classifier based on Estimation of Supporters: uses as features the PageRank, the estimation of supporters at a given distance $d = 2, 3, 4$, and the estimation of supporters divided by PageRank.
### Comparison of single-technique classifier (M=5)

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Spam class</th>
<th>False Pos.</th>
<th>False Neg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(pruning with $M = 5$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TrustRank</td>
<td>0.82 0.50 2.1% 50%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trunc. PageRank $t = 2$</td>
<td>0.85 0.50 1.6% 50%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trunc. PageRank $t = 3$</td>
<td>0.84 0.47 1.6% 53%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trunc. PageRank $t = 4$</td>
<td>0.79 0.45 2.2% 55%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Est. Supporters $d = 2$</td>
<td>0.78 0.60 3.2% 40%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Est. Supporters $d = 3$</td>
<td>0.83 0.64 2.4% 36%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Est. Supporters $d = 4$</strong></td>
<td><strong>0.86 0.64 2.0% 36%</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Comparison of single-technique classifier (M=30)

<table>
<thead>
<tr>
<th>Classifiers (pruning with $M = 30$)</th>
<th>Spam Prec.</th>
<th>Spam Recall</th>
<th>False Pos.</th>
<th>False Neg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TrustRank</td>
<td>0.80</td>
<td>0.49</td>
<td>2.3%</td>
<td>51%</td>
</tr>
<tr>
<td>Trunc. PageRank $t = 2$</td>
<td>0.82</td>
<td>0.43</td>
<td>1.8%</td>
<td>57%</td>
</tr>
<tr>
<td>Trunc. PageRank $t = 3$</td>
<td>0.81</td>
<td>0.42</td>
<td>1.8%</td>
<td>58%</td>
</tr>
<tr>
<td>Trunc. PageRank $t = 4$</td>
<td>0.77</td>
<td>0.43</td>
<td>2.4%</td>
<td>57%</td>
</tr>
<tr>
<td>Est. Supporters $d = 2$</td>
<td>0.76</td>
<td>0.52</td>
<td>3.1%</td>
<td>48%</td>
</tr>
<tr>
<td><strong>Est. Supporters $d = 3$</strong></td>
<td><strong>0.83</strong></td>
<td><strong>0.57</strong></td>
<td><strong>2.1%</strong></td>
<td><strong>43%</strong></td>
</tr>
<tr>
<td>Est. Supporters $d = 4$</td>
<td>0.80</td>
<td>0.57</td>
<td>2.6%</td>
<td>43%</td>
</tr>
</tbody>
</table>
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

<table>
<thead>
<tr>
<th>Pruning</th>
<th>Rules</th>
<th>Precision</th>
<th>Recall</th>
<th>False Pos.</th>
<th>False Neg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>M=5</td>
<td>49</td>
<td>0.87</td>
<td>0.80</td>
<td>2.0%</td>
<td>20%</td>
</tr>
<tr>
<td>M=30</td>
<td>31</td>
<td>0.88</td>
<td>0.76</td>
<td>1.8%</td>
<td>24%</td>
</tr>
<tr>
<td>No pruning</td>
<td>189</td>
<td>0.85</td>
<td>0.79</td>
<td>2.6%</td>
<td>21%</td>
</tr>
</tbody>
</table>
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

1 Motivation
2 Spam pages characterization
3 Truncated PageRank
4 Counting supporters
5 Experiments
6 Conclusions
Summary of classifiers

(a) Precision and recall of spam detection

(b) Error rates of the spam classifiers
Conclusions

Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation

Spam pages characterization

Truncated PageRank

Counting supporters

Experiments

Conclusions

✔️ Link-based statistics to detect 80% of spam
Conclusions

- Link-based statistics to detect 80% of spam
- No magic bullet in link analysis
Conclusions

✔️ Link-based statistics to detect 80% of spam
❌ No magic bullet in link analysis
❌ Precision still low compared to e-mail spam filters
Conclusions

- Link-based statistics to detect 80% of spam
- No magic bullet in link analysis
- Precision still low compared to e-mail spam filters
- Measure both home page and max. PageRank page
Conclusions

- Link-based statistics to detect 80% of spam
- No magic bullet in link analysis
- Precision still low compared to e-mail spam filters
- Measure both home page and max. PageRank page
- Host-based counts are important

Motivation
- Spam pages characterization
- Truncated PageRank
- Counting supporters
- Experiments
- Conclusions
Conclusions

- Link-based statistics to detect 80% of spam
- No magic bullet in link analysis
- Precision still low compared to e-mail spam filters
- Measure both home page and max. PageRank page
- Host-based counts are important
Conclusions

- Link-based statistics to detect 80% of spam
- No magic bullet in link analysis
- Precision still low compared to e-mail spam filters
- Measure both home page and max. PageRank page
- Host-based counts are important

Next step: combine link analysis and content analysis
Thank you!


Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation
Spam pages characterization
Truncated PageRank
Counting supporters
Experiments
Conclusions

Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages.
In Proceedings of the seventh workshop on the Web and databases (WebDB), pages 1–6, Paris, France.

Probabilistic counting algorithms for data base applications.

Discovering large dense subgraphs in massive graphs.
Using rank propagation and Probabilistic counting for Link-Based Spam Detection

L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation
Spam pages characterization
Truncated PageRank
Counting supporters
Experiments
Conclusions

Web spam taxonomy.
In First International Workshop on Adversarial Information Retrieval on the Web.

Combating web spam with trustrank.
In Proceedings of the Thirtieth International Conference on Very Large Data Bases (VLDB), pages 576–587, Toronto, Canada. Morgan Kaufmann.

Random graphs with arbitrary degree distributions and their applications.
