Link-Based Spam Detection

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

Motivation

Degree-based measures

PageRank

TrustRank

Truncated PageRank

Counting supporters

Conclusions

Using Rank Propagation and Probabilistic Counting for Link-based Spam Detection

Luca Becchetti[1], Carlos Castillo[1],Debora Donato[1], Stefano Leonardi[1] and Ricardo Baeza-Yates[2]

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

May 19th, 2006

Link-Based Spam Detection

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

Motivation

Degree-based measures

PageRank

TrustRank

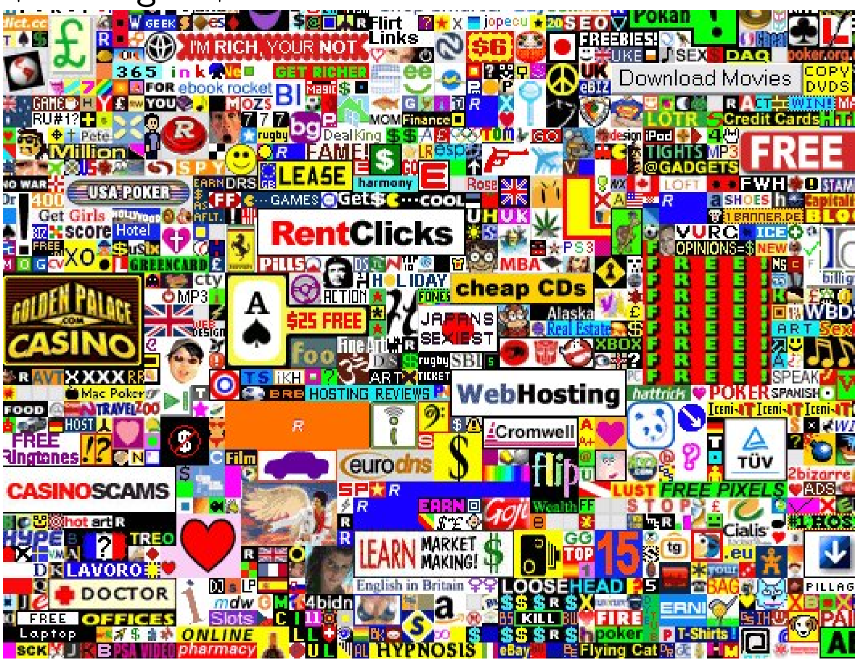
Truncated PageRank

Counting supporters

Conclusions

What is on the Web?

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



Graphic: www.milliondollarhomepage.com

Web spam (keywords + links)

viagra buy viagra viagradrugs.net, to cialis lawsuit, dirt cheap viagra, in sex discount cialis generic cialis bluepilled.com, herbal alternative viagra, for cialis marijuana, sublingual viagra.
Viagra users, will viagra facts cialis line prescription, buy viagra online viagra side effects natural alternative viagra, has cialis generic viagra generic cialis cialis cum-with-us.com, viagra discount, this brand name cialis, herbal viagra alternative free viagra buying deal viagradrugs.net cheapest price viagra cheap viagra uk free viagra viagra online pills pills viagradrugs.net, silagra weight loss generic viagra cialis cum-with-us.com, viagra blindness viagra prescription.
Amsterdam viagra sexshops viagra prescription for woman viagra online pharmacy, is cialis ordering online, viagra suppliers cocaine and viagra sex experiences viagra generico impotencia, cialis official website, viagra cheap generic cheap viagra natural viagra, will ciali, whats the chemical name for the drug viagra, are cialis and grapefruit, homemade viagra, has herbal cialis, strength of erection viagra levitra cialis.

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

- generic viagra
- buy viagra
- viagra alternative
- herbal viagra
- cheap viagra
- viagra online
- buy viagra online
- order viagra
- order viagra online
- Viagra
- natural viagra
- viagra pill
- free viagra samples
- discount viagra
- female viagra
- viagra

Web spam (mostly keywords)

Link-Based Spam Detection

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

Motivation

Degree-based measures

PageRank

TrustRank

Truncated PageRank

Counting supporters

Conclusions

smart movie converter 2.72 registration key as nokia6600

crac diablo2, kontakt 2 .exe, download crack norton 2006 online, dowloand snagit, telecharger canopu nothing else mather/ lyric, silent hill 3 no-dvd crack, Protocol v7 Update-Vengeance.rar download, kn windows validation crack download

donwload counter strike 1 5

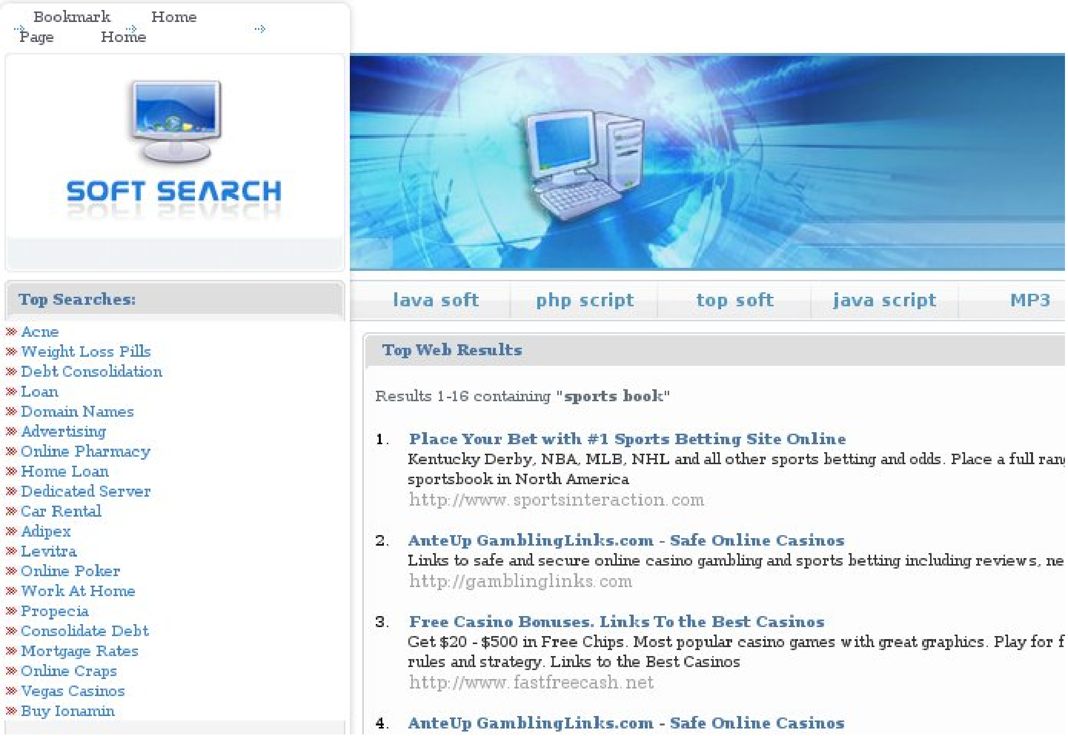
ftp downloads spanish, total commander hack, plus superpack key, crack de need ford speed underground, manual diablo 2, Rapture_dhol_mix.mp3, advance 3gp convertor, pacific assault torrent, Fruity Loops 4.5.1 Demo, russian mohaa skin, telechargement de crack battlefield 2

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, downlod 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 cracks, winamp crack licens, mobil msn sis, nero burning room 6.0.0.19, Deep Silver Keygen, the sims makin ` magik exploration pack Flash Get, DesktopX) 3.1 (keygen | serial, application architecture control, download ogc quake3

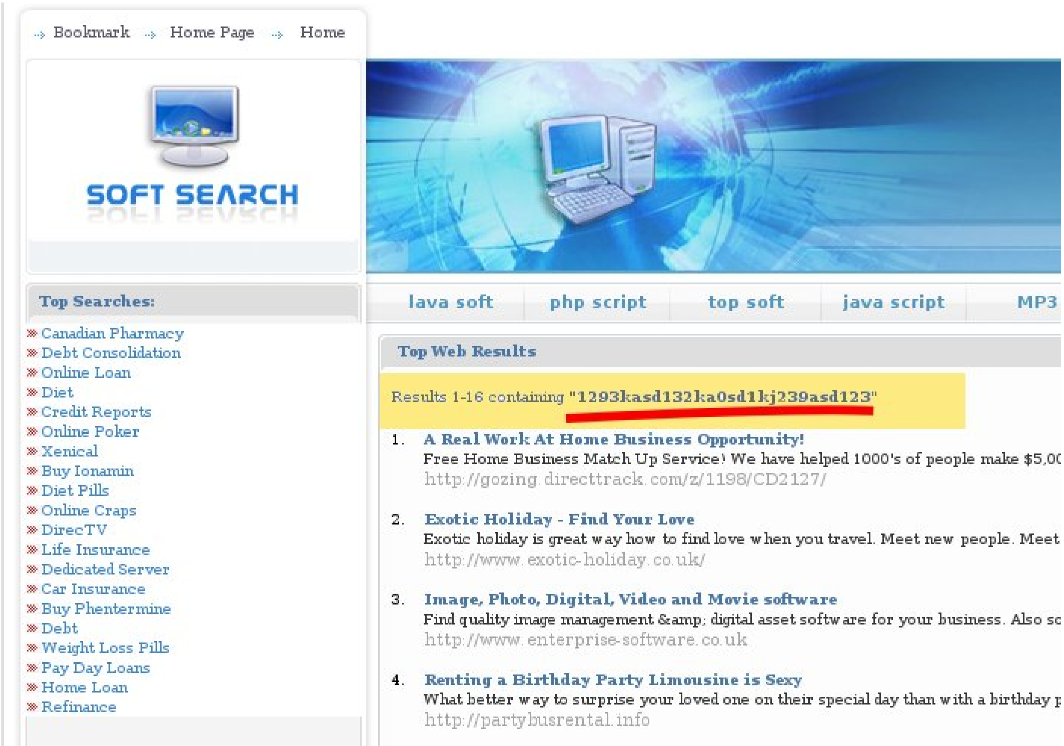
free gemu, telecharger cdkeys, care day home question, Remote S60 software cracked, nt print server, SWF2Video Plugin for Adobe Premiere Pro cracked, ps2 secret code **download webcam lv-c300,**

Search engine?

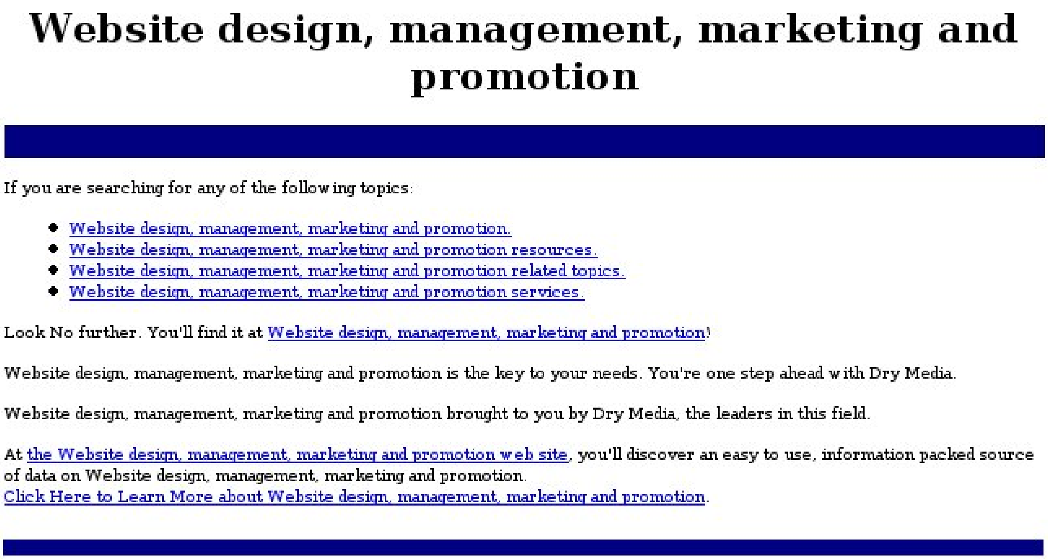


Fake search engine

Link-Based Spam Detection
L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates
Motivation
Degree-based measures
PageRank
TrustRank
Truncated PageRank
Counting supporters
Conclusions



Problem: "normal" pages that are spam



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!

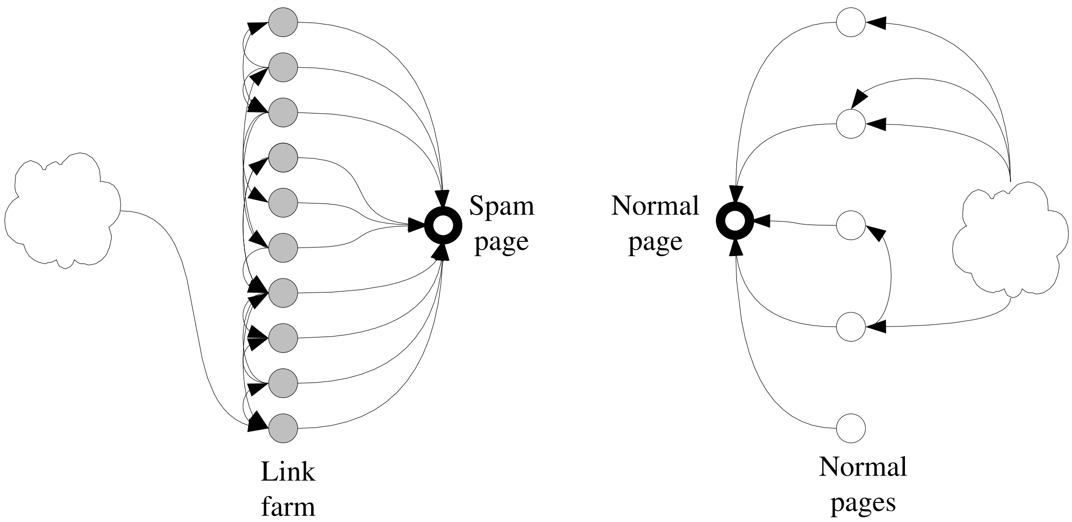
Website design, management, marketing and promotion is the key to your needs. You're one step ahead with Dry Media.

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.

Link farms

Link-Based Spam Detection
L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates
Motivation
Degree-based measures
PageRank
TrustRank
Truncated PageRank
Counting supporters
Conclusions



Single-level farms can be detected by searching groups of nodes sharing their out-links [Gibson et al., 2005]

Motivation

[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

Metrics

Link-Based Spam Detection
L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates
Motivation
Degree-based measures
PageRank
TrustRank
Truncated PageRank
Counting supporters
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

Test collection

U.K. collection

18.5 million pages downloaded from the .UK domain

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

Classified entire hosts:

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

Link-Based Spam Detection

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

Motivation

Degree-based measures

PageRank

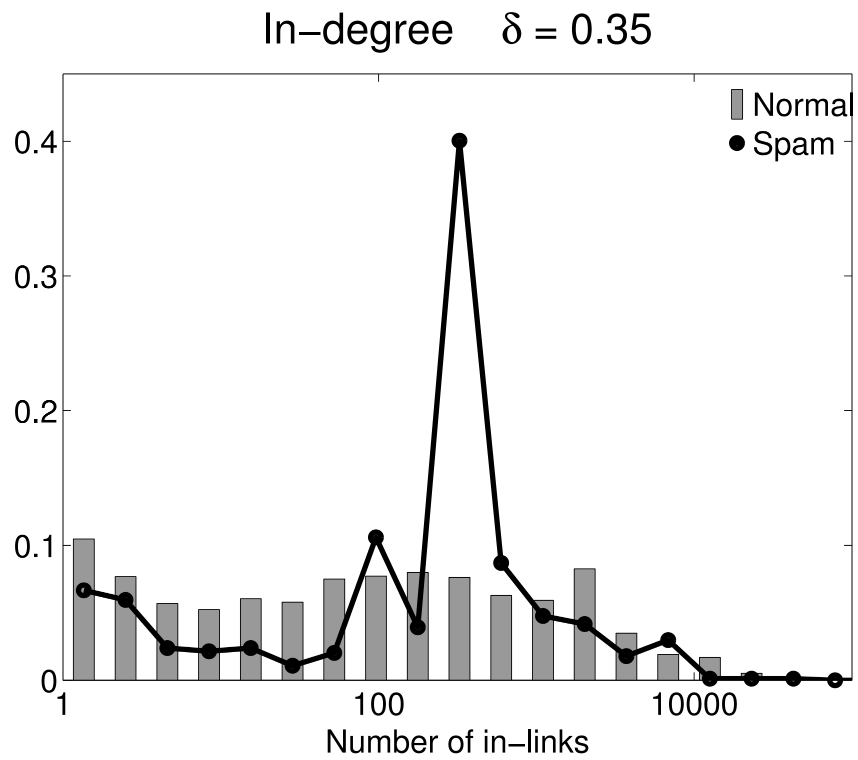
TrustRank

Truncated PageRank

Counting supporters

Conclusions

Degree



($\delta =$ max. difference in C.D.F. plot)

Degree

Link-Based
Spam Detection

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

Motivation

Degree-based
measures

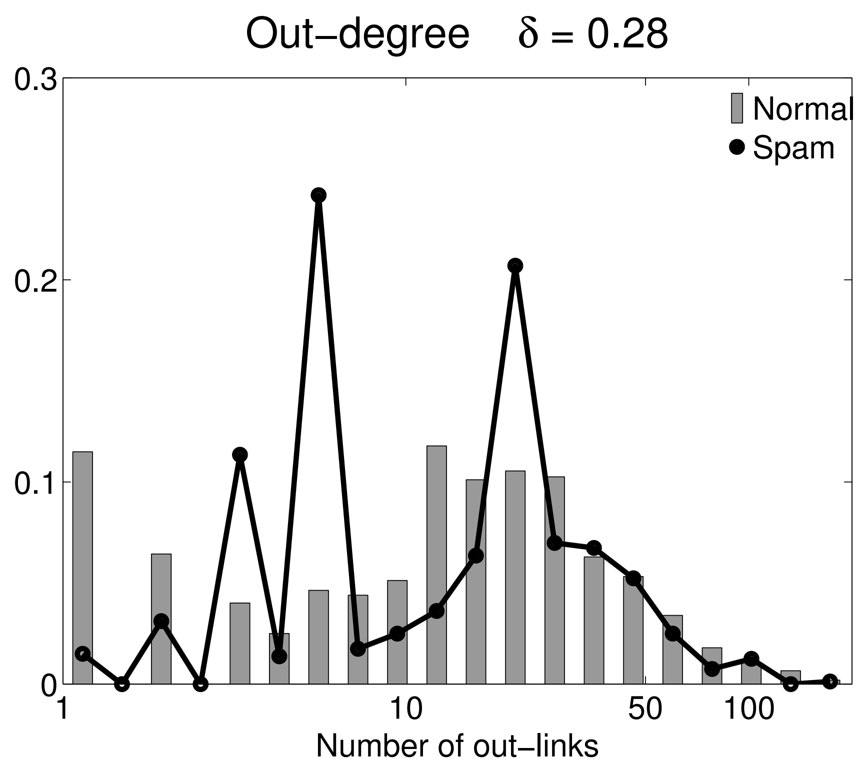
PageRank

TrustRank

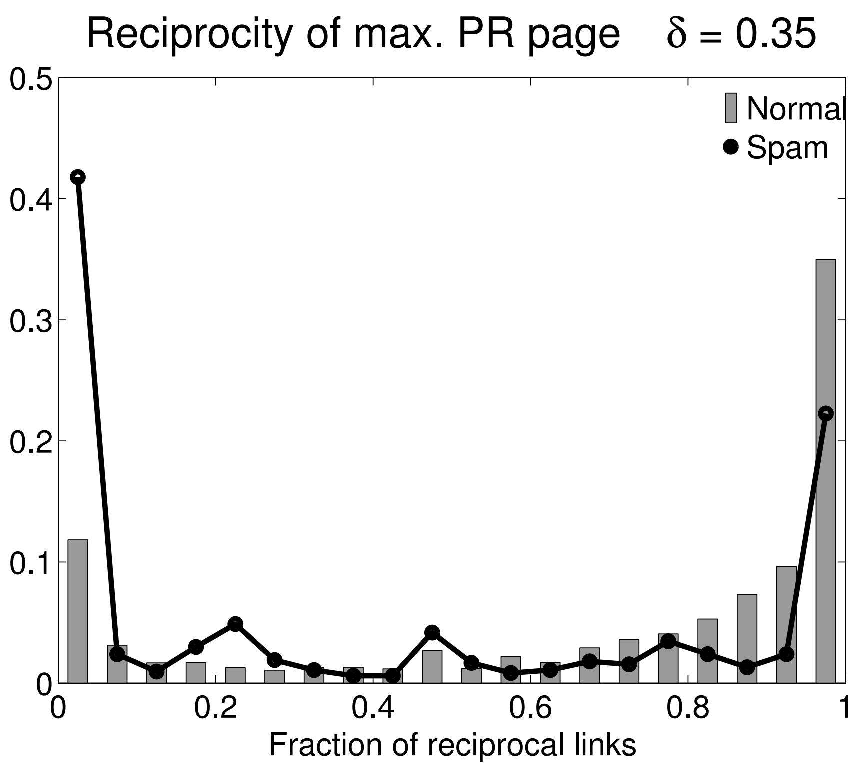
Truncated
PageRank

Counting
supporters

Conclusions

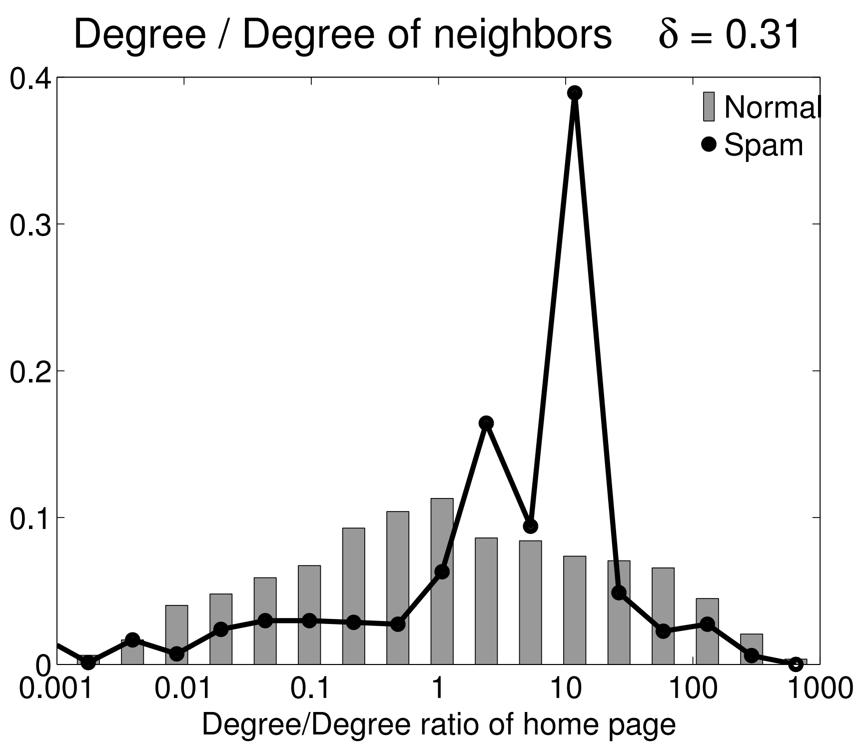


Edge reciprocity



Assortativity

Link-Based Spam Detection
L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates
Motivation
Degree-based measures
PageRank
TrustRank
Truncated PageRank
Counting supporters
Conclusions



Automatic classifier

All of the following attributes in the home page and the page with the maximum PageRank, plus a binary variable indicating if they are the same page:

- In-degree, out-degree
- Fraction of reciprocal edges
- Degree divided by degree of direct neighbors
- Average and sum of in-degree of out-neighbors
- Average and sum of out-degree of in-neighbors

Decision tree

72.6% of detection rate, with 3.1% false positives

Link-Based Spam Detection

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

Motivation

Degree-based measures

PageRank

TrustRank

Truncated PageRank

Counting supporters

Conclusions

PageRank

Let $\mathbf{P}_{N \times N}$ be the normalized link matrix of a graph

- Row-normalized
- No "sinks"

Definition (PageRank)

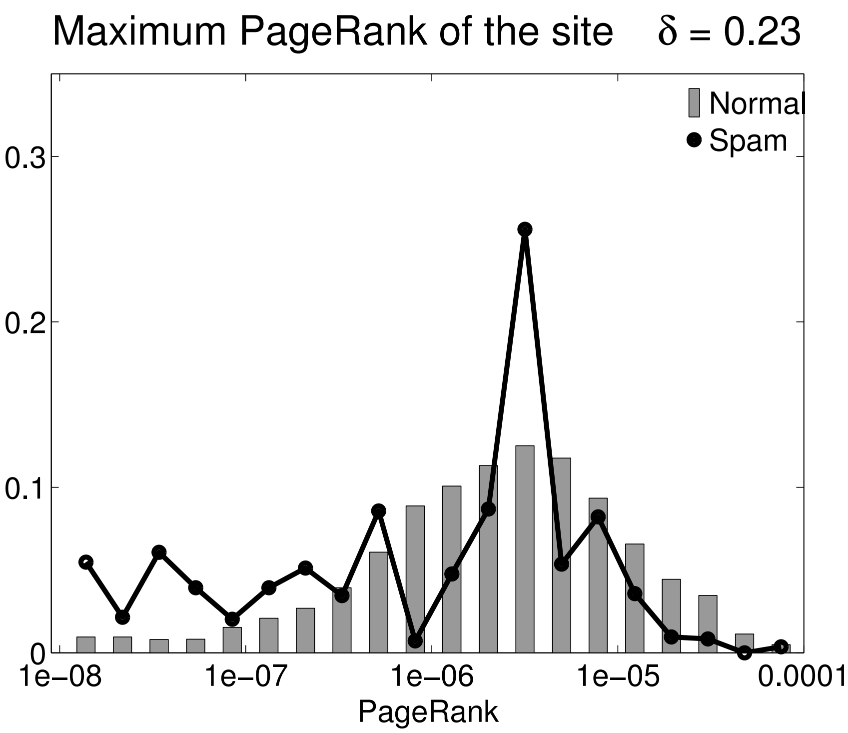
Stationary state of:

$$\alpha \mathbf{P} + \frac{(1-\alpha)}{N} \mathbf{1}_{N \times N}$$

- Follow links with probability α
- Random jump with probability $1 - \alpha$

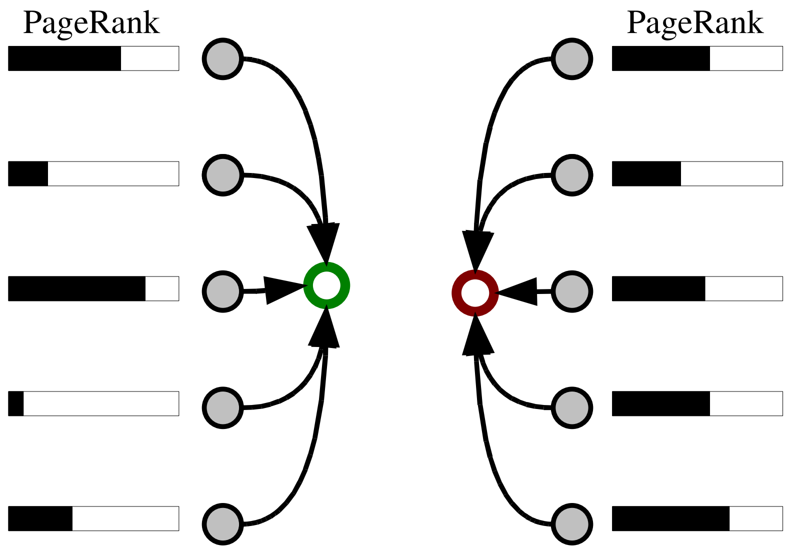
Link-Based Spam Detection
L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates
Motivation
Degree-based measures
PageRank
TrustRank
Truncated PageRank
Counting supporters
Conclusions

Maximum PageRank in the Host



Variance of PageRank

Suggested in [Benczúr et al., 2005]

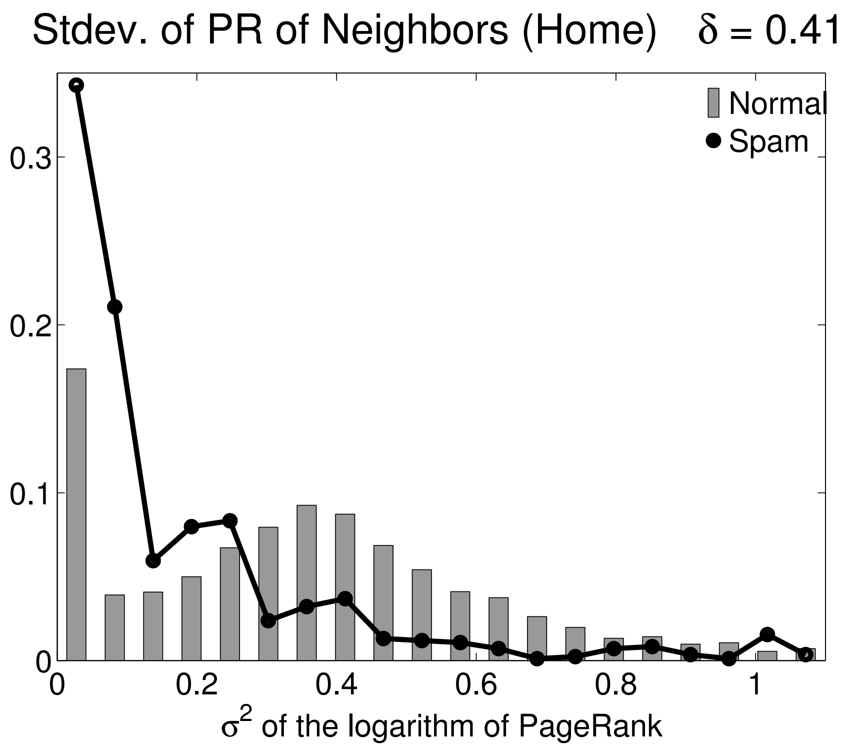


Variance of PageRank of in-neighbors

Link-Based Spam Detection

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

Motivation
Degree-based measures
PageRank
TrustRank
Truncated PageRank
Counting supporters
Conclusions



Automatic classifier

Features: degree-based plus the following in the home page and the page with maximum PageRank:

- PageRank
- In-degree/PageRank
- Out-degree/PageRank
- Standard deviation of PageRank of in-neighbors $= \sigma^2$
- σ^2/PageRank

Plus the PageRank of the home page divided by the PageRank of the page with the maximum PageRank.

Decision tree

74.4% of detection rate, with 2.6% false positives
(Degree-based: 72.6% of detection, 3.1% false positives)

Link-Based Spam Detection

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

Motivation
Degree-based measures
PageRank
TrustRank
Truncated PageRank
Counting supporters
Conclusions

TrustRank

TrustRank [Gyöngyi et al., 2004]

A node with high PageRank, but far away from a core set of "trusted nodes" is suspicious

Start from a set of trusted nodes, then do a random walk, returning to the set of trusted nodes with probability $1 - \alpha$ at each step

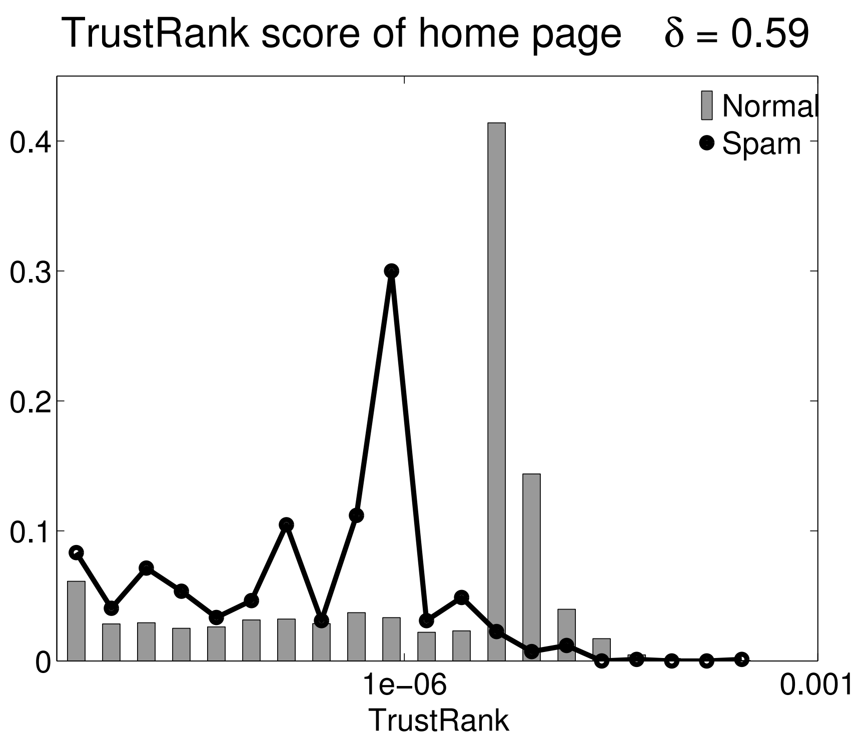
Trusted nodes: data from `http://www.dmoz.org/`

TrustRank score

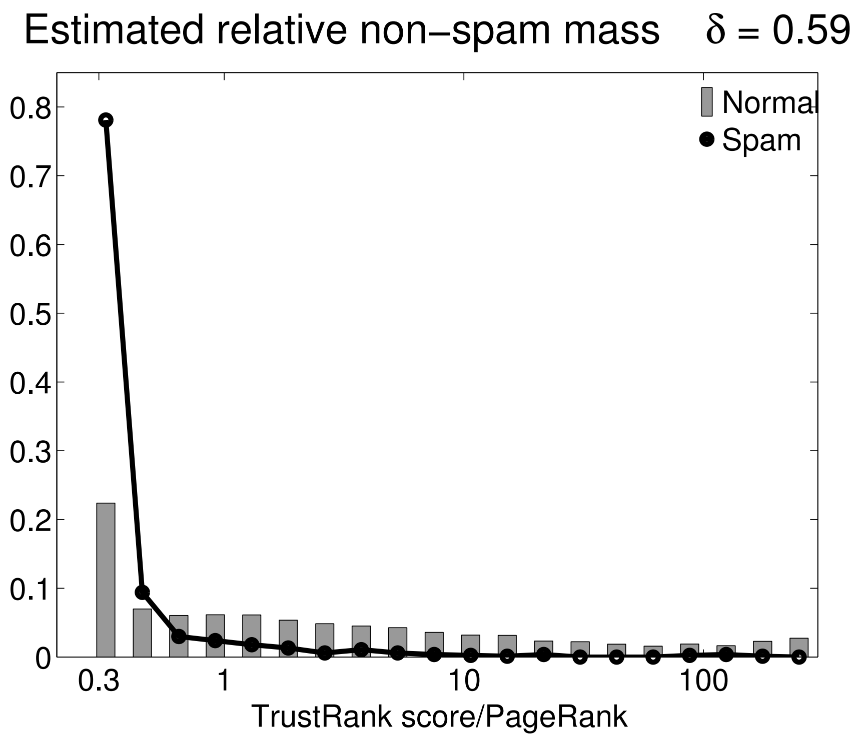
Link-Based
Spam Detection

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

Motivation
Degree-based measures
PageRank
TrustRank
Truncated PageRank
Counting supporters
Conclusions



TrustRank / PageRank



Link-Based
Spam Detection

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

Motivation
Degree-based measures
PageRank
TrustRank
Truncated PageRank
Counting supporters
Conclusions

Automatic classifier

PageRank attributes, plus the following in the home page and the page with maximum PageRank:

- TrustScore
- TrustScore/PageRank (estimated relative non-spam mass)
- TrustScore/In-degree

Plus the TrustScore in the home page divided by the TrustScore in the page with the maximum PageRank.

Decision tree

77.3% of detection rate, with 3.0% false positives
(PageRank-based: 74.4% of detection, 2.6% false positives)

Path-based formula for PageRank

Link-Based Spam Detection

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

Motivation

Degree-based measures

PageRank

TrustRank

Truncated PageRank

Counting supporters

Conclusions

Given a path $p = \langle x_1, x_2, \ldots, x_t \rangle$ of length $t = |p|$

$$\text{branching}(p) = \frac{1}{d_1 d_2 \cdots d_{t-1}}$$

where d_i are the out-degrees of the members of the path

Explicit formula for PageRank [Newman et al., 2001]

$$r_i(\alpha) = \sum_{p \in \text{Path}(-,i)} \frac{(1-\alpha)\alpha^{|p|}}{N} \text{branching}(p)$$

$\text{Path}(-, i)$ are incoming paths in node i

General functional ranking

In general:

$$r_i(\alpha) = \sum_{p \in \mathrm{Path}(-,i)} \frac{\mathrm{damping}(|p|)}{N} \mathrm{branching}(p)$$

There are many choices for damping($|p|$), including simply a linear function that is as good as PageRank in practice [Baeza-Yates et al., 2006]

Link-Based Spam Detection

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

Motivation

Degree-based measures

PageRank

TrustRank

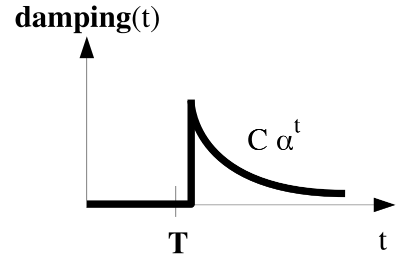
Truncated PageRank

Counting supporters

Conclusions

Truncated PageRank

Reduce the direct contribution of the first levels of links:



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

☑ No extra reading of the graph after PageRank

Truncated PageRank(T=2) / PageRank

Link-Based
Spam Detection

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

Motivation

Degree-based measures

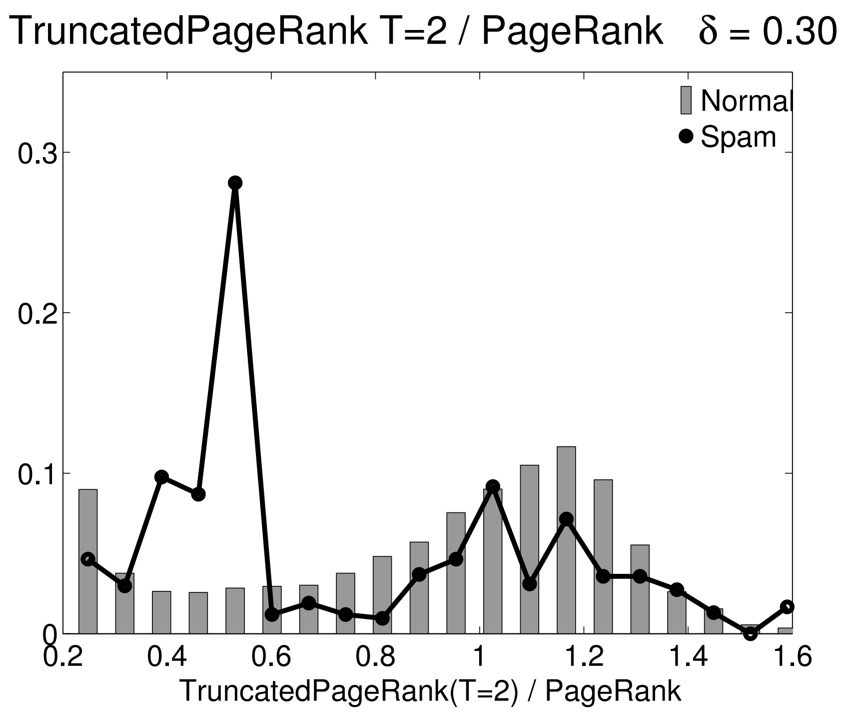
PageRank

TrustRank

Truncated PageRank

Counting supporters

Conclusions



Max. change of Truncated PageRank

Link-Based
Spam Detection

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

Motivation

Degree-based measures

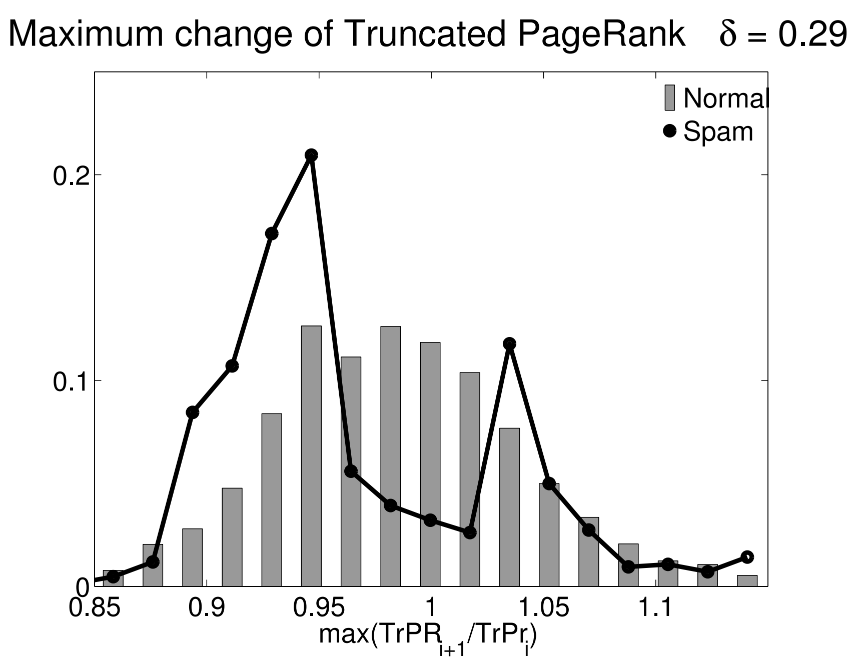
PageRank

TrustRank

Truncated PageRank

Counting supporters

Conclusions



Automatic classifier

PageRank attributes, plus the following in the home page and the page with maximum PageRank:

- TruncPageRank($T = 1 \ldots 4$)
- TruncPageRank($T = 4$) / TruncPageRank($T = 3$)
- TruncPageRank($T = 3$) / TruncPageRank($T = 2$)
- TruncPageRank($T = 2$) / TruncPageRank($T = 1$)
- TruncPageRank($T = 1 \ldots 4$) / PageRank
- Minimum, maximum and average of:

TruncPageRank($T = i$)/TruncPageRank($T = i - 1$)

Plus the TruncatedPageRank($T = 1 \ldots 4$) of the home page divided by the same value in the page with the maximum PageRank.

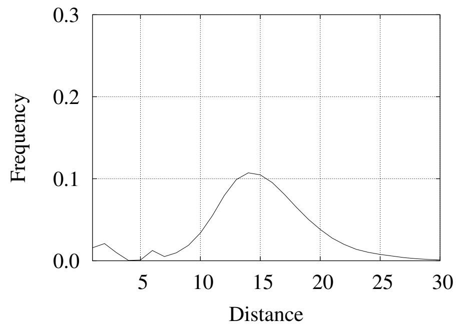
Decision tree

76.9% of detection rate, with 2.5% false positives
(TrustRank-based: 77.3% of detection, 3.0% false positives)

Idea: count "supporters" at different distances

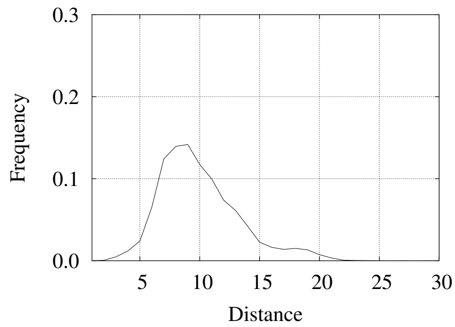
Number of different nodes at a given distance:

.UK 18 mill. nodes



Average distance
14.9 clicks

.EU.INT 860,000 nodes



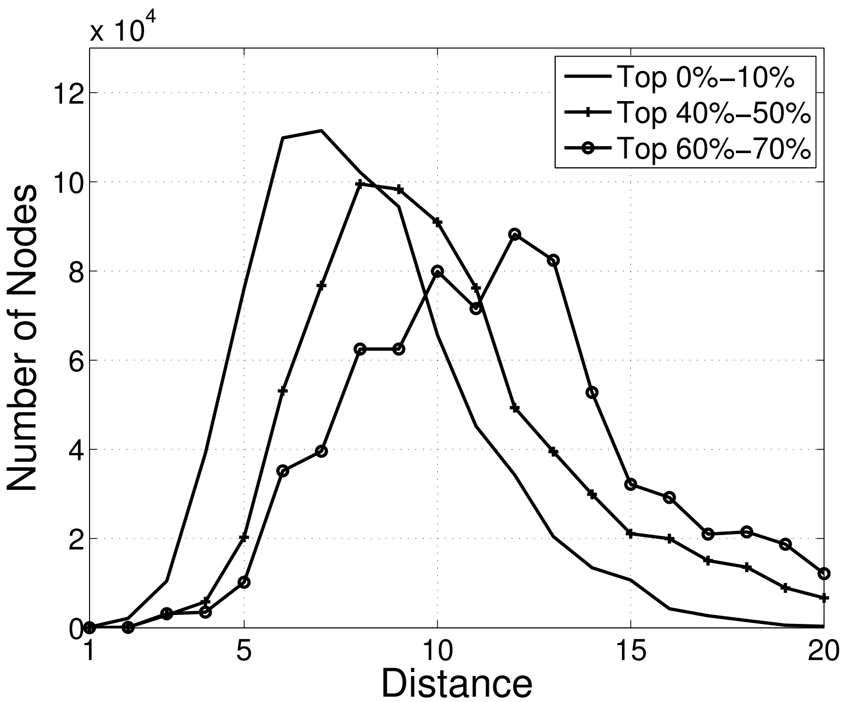
Average distance
10.0 clicks

High and low-ranked pages are different

Link-Based Spam Detection

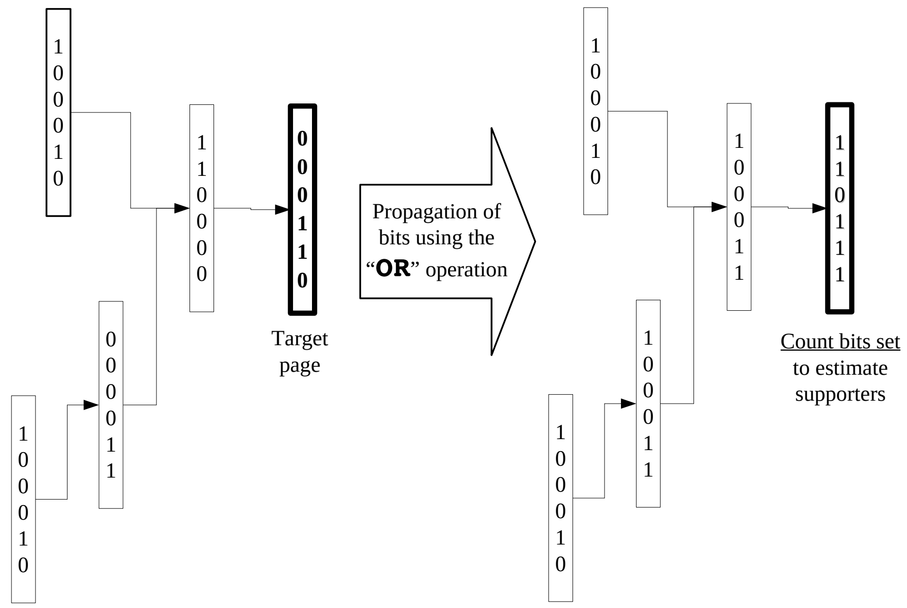
L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates

Motivation
Degree-based measures
PageRank
TrustRank
Truncated PageRank
Counting supporters
Conclusions



Areas below the curves are equal if we are in the same strongly-connected component

Probabilistic counting



Improvement of ANF algorithm [Palmer et al., 2002] based on probabilistic counting [Flajolet and Martin, 1985]

General algorithm

Link-Based Spam Detection

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

Motivation
Degree-based measures
PageRank
TrustRank
Truncated PageRank
Counting supporters
Conclusions

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

for node : 1 . . . N, bit: 1 . . . k **do**
    `INIT`(node,bit)
end for
for distance : 1 . . . d **do** {Iteration step}
    Aux $\leftarrow \mathbf{0}_k$
    for src : 1 . . . N **do** {Follow links in the graph}
        for all links from src to dest **do**
            Aux[dest] $\leftarrow$ Aux[dest] OR V[src,$\cdot$]
        end for
    end for
    V $\leftarrow$ Aux
end for
for node: 1 . . . N **do** {Estimate supporters}
    Supporters[node] $\leftarrow$ `ESTIMATE`(V[node,$\cdot$])
end for
return Supporters

Our estimator

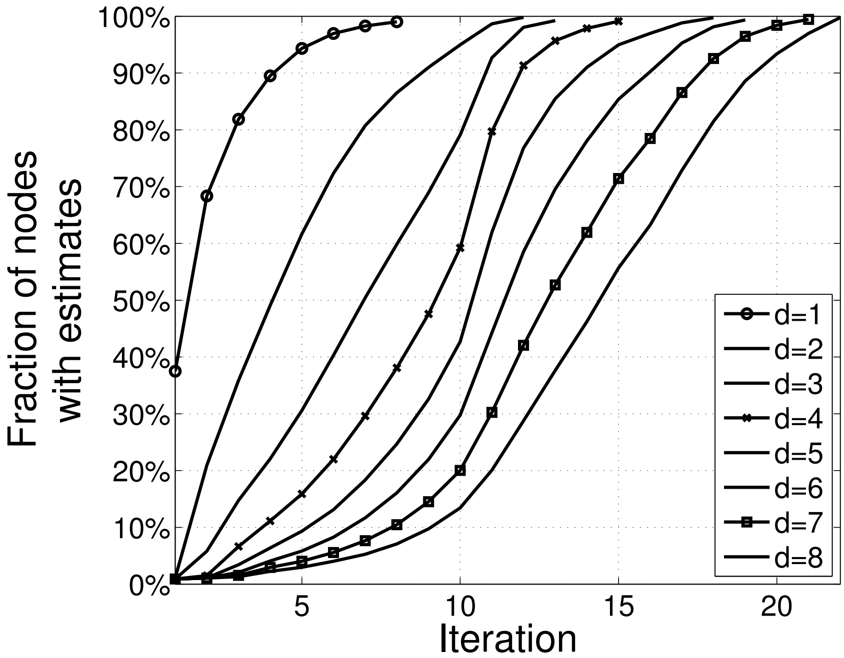
Initialize all bits to one with probability ϵ

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

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.

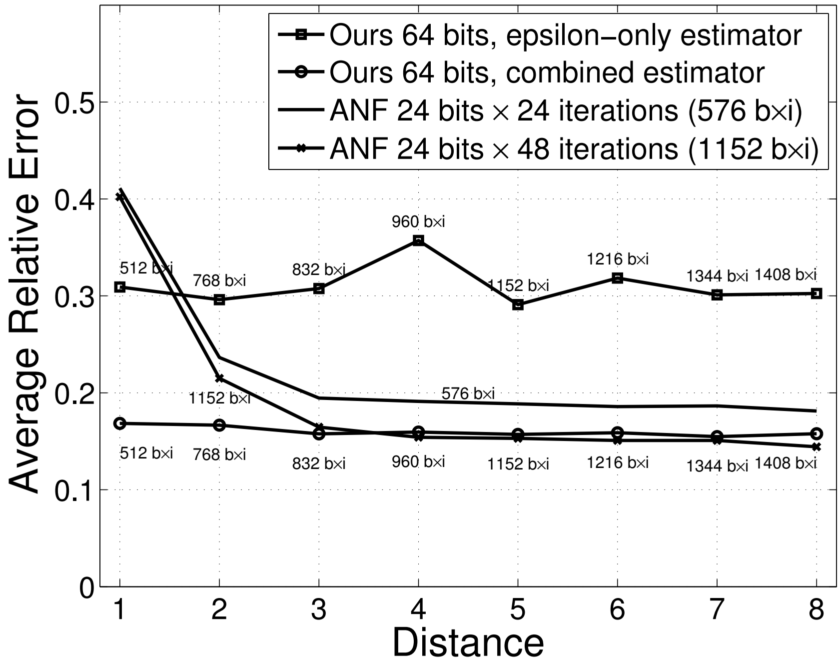
Convergence



Error rate

Link-Based Spam Detection

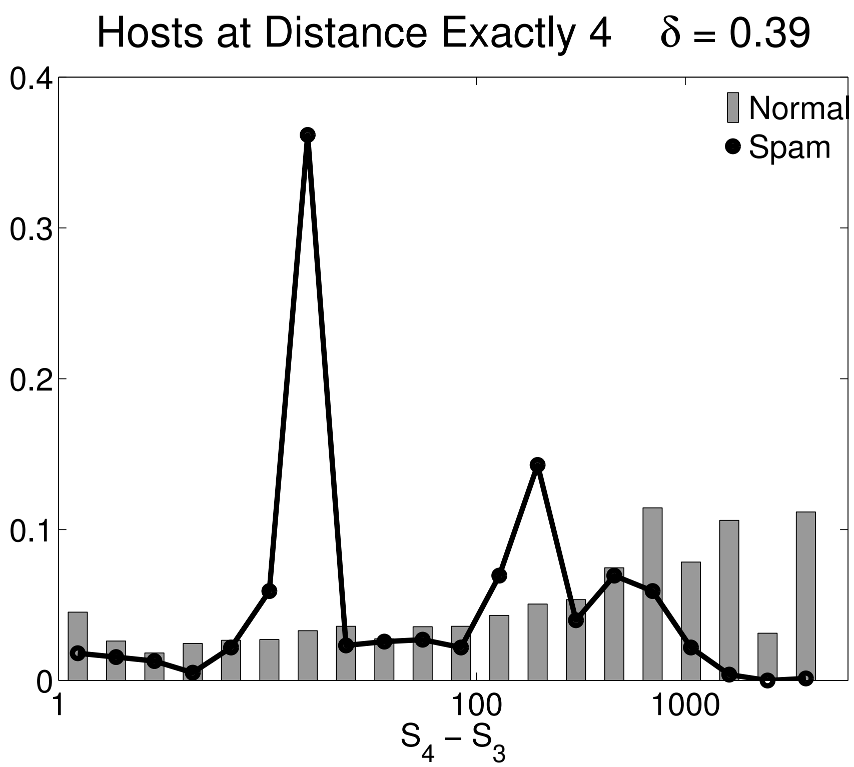
L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yates



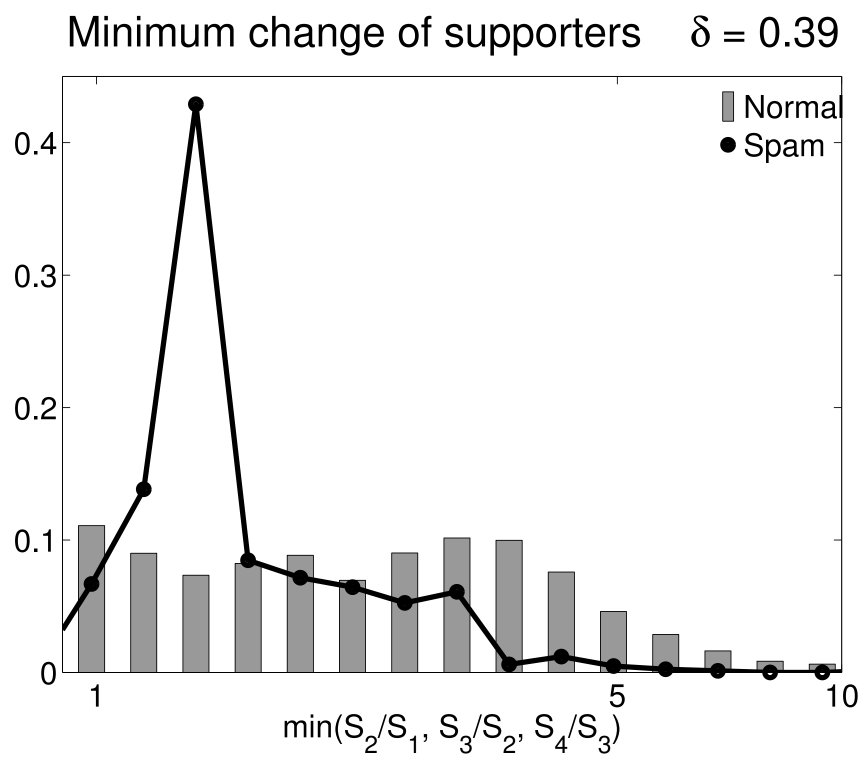
Hosts at distance 4

Link-Based Spam Detection

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



Minimum change of supporters



Automatic classifier

Link-Based
Spam Detection

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

Motivation
Degree-based measures
PageRank
TrustRank
Truncated PageRank
Counting supporters
Conclusions

PageRank attributes, plus the following in the home page and the page with maximum PageRank:

- Supporters at $2 \ldots 4$
- Supporters at $2 \ldots 4$ / PageRank
- Supporters at i / Supporters at $i-1$ (for $i = 1..4$)
- Minimum, maximum and average of: Supporters at i / Supporters at $i-1$ (for $i = 1..4$)
- (Supporters at i - Supporters at $i-1$) / PageRank

Plus the number of supporters at distance $2 \ldots 4$ in the home page divided by the same feature in the page with the maximum PageRank.

Decision tree

78.9% of detection rate, with 2.5% false positives
(TruncPR: 76.9% of detection, 2.5% false positives)
(TrustRank: 77.7% of detection, 3.0% false positives)

Summary of classifiers

Link-Based
Spam Detection

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

Motivation

Degree-based measures

PageRank

TrustRank

Truncated PageRank

Counting supporters

Conclusions

Metrics	Detection rate	False positives
Degree (D)	72.6%	3.1%
D + PageRank (P)	74.4%	2.6%
D + P + TrustRank	77.7%	3.0%
D + P + Trunc. PageRank	76.9%	2.5%
D + P + Est. of Supporters	78.9%	2.5%
All attributes	81.4%	2.8%
All attributes (more rules)	80.8%	1.1%

Comparison

Content-based analysis [Ntoulas et al., 2006] has shown 86.2% detection rate with 2.2% false positives

Single-attribute classifier with TrustRank [Gyöngyi et al., 2004]: 51.1% detection rate, 3.4% error in our sample – SpamRank [Benczúr et al., 2005] reports similar detection rates

Top 10 metrics

1. Binary variable indicating if homepage is the page with maximum PageRank of the site
2. Edge reciprocity
3. Different hosts at distance 4
4. Different hosts at distance 3
5. Minimum change of supporters (different hosts)
6. Different hosts at distance 2
7. TruncatedPagerank (T=1) / PageRank
8. TrustRank score divided by PageRank
9. Different hosts at distance 1
10. TruncatedPagerank (T=2) / PageRank

Conclusions

Link-Based Spam Detection

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

Motivation
Degree-based measures
PageRank
TrustRank
Truncated PageRank
Counting supporters
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!

Link-Based
Spam Detection

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

Motivation

Degree-based
measures

PageRank

TrustRank

Truncated
PageRank

Counting
supporters

Conclusions

Baeza-Yates, R., Boldi, P., and Castillo, C. (2006).
Generalizing pagerank: Damping functions for link-based ranking algorithms.
In *Proceedings of SIGIR*, Seattle, Washington, USA. ACM Press.

Becchetti, L., Castillo, C., Donato, D., Leonardi, S., and Baeza-Yates, R. (2006).
Using rank propagation and probabilistic counting for link-based spam detection.
Technical report, DELIS – Dynamically Evolving, Large-Scale Information Systems.

Benczúr, A. A., Csalogány, K., Sarlós, T., and Uher, M. (2005).
Spamrank: fully automatic link spam detection.
In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, Chiba, Japan.

Fetterly, D., Manasse, M., and Najork, M. (2004).
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.

Flajolet, P. and Martin, N. G. (1985).
Probabilistic counting algorithms for data base applications.
Journal of Computer and System Sciences, 31(2):182–209.

Gibson, D., Kumar, R., and Tomkins, A. (2005).
Discovering large dense subgraphs in massive graphs.
In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pages 721–732. VLDB Endowment.

Link-Based Spam Detection

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

Motivation

Degree-based measures

PageRank

TrustRank

Truncated PageRank

Counting supporters

Conclusions

Gyöngyi, Z., Molina, H. G., and Pedersen, J. (2004).
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.

Newman, M. E., Strogatz, S. H., and Watts, D. J. (2001).
Random graphs with arbitrary degree distributions and their applications.
Phys Rev E Stat Nonlin Soft Matter Phys, 64(2 Pt 2).

Ntoulas, A., Najork, M., and Manasse, M. a. (2006).
Detecting spam web pages through content analysis.
In *To appear in proceedings of the World Wide Web conference*, Edinburgh, Scotland.

Palmer, C. R., Gibbons, P. B., and Faloutsos, C. (2002).

ANF: a fast and scalable tool for data mining in massive graphs.

In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 81–90, New York, NY, USA. ACM Press.