

# Web Spam Detection

**Ricardo Baeza-Yates**<sup>1,3</sup>

ricardo@baeza.cl

With: L. Becchetti<sup>2</sup>, P. Boldi<sup>5</sup>, C. Castillo<sup>1</sup>, D. Donato<sup>1</sup>,  
A. Gionis<sup>1</sup>, S. Leonardi<sup>2</sup>, V. Murdock<sup>1</sup>, M. Santini<sup>5</sup>,  
F. Silvestri<sup>4</sup>, S. Vigna<sup>5</sup>

1. Yahoo! Research Barcelona – Catalunya, Spain

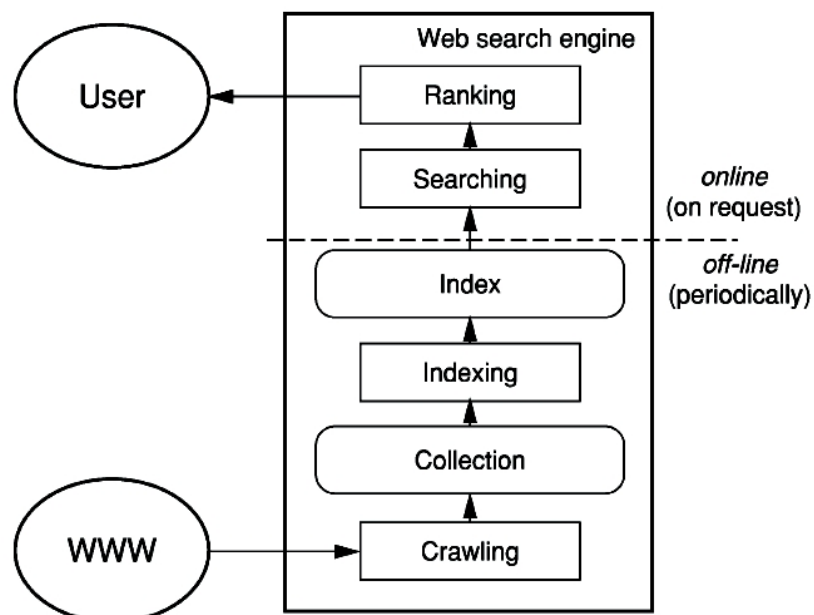
2. Università di Roma “La Sapienza” – Rome, Italy

3. Yahoo! Research Santiago – Chile

4. ISTI-CNR –Pisa, Italy

5. Università degli Studi di Milano – Milan, Italy

## Previous: how search engines work

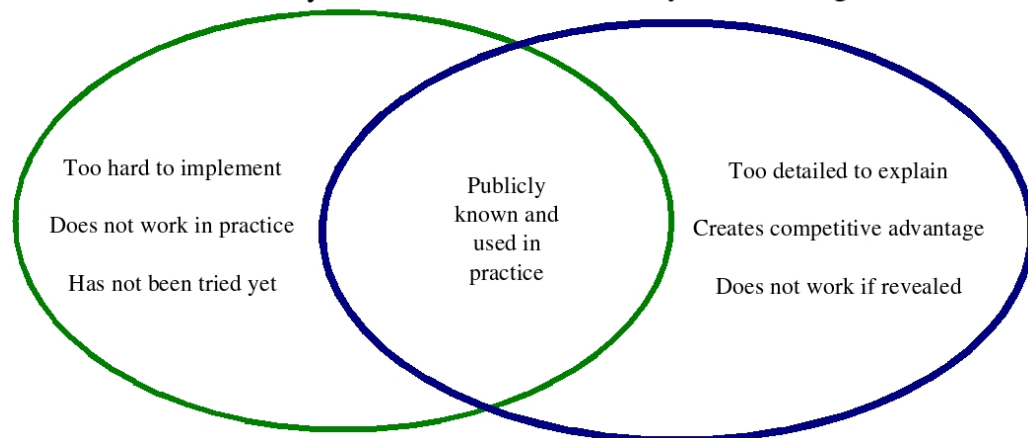


# Search engine: issues

- Scalability (crawling, indexing, searching, ranking)
- Relevance (query to document match)
- Static ranking (content quality)
- Incentives for cheating (\$)

# This is a talk about academic research!

## Tools for dealing with Web Spam



- 1 Web Spam
- 2 Web Spam Detection
- 3 A Reference Collection
- 4 Web Links
- 5 Topological Web Spam
- 6 Counting of Supporters
- 7 Content-based Spam detection
- 8 Web Topology
- 9 Conclusions

- 1 Web Spam
- 2 Web Spam Detection
- 3 A Reference Collection
- 4 Web Links
- 5 Topological Web Spam
- 6 Counting of Supporters
- 7 Content-based Spam detection
- 8 Web Topology
- 9 Conclusions

# The Web

“The sum of all human knowledge plus porn” – Robert Gilbert



Graphic: [www.milliondollarhomepage.com](http://www.milliondollarhomepage.com)

# Adversarial IR Issues on the Web

- Link spam
- Content spam
- Cloaking
- Comment/forum/wiki spam
- Spam-oriented blogging
- Click fraud  $\times 2$
- Reverse engineering of ranking algorithms
- Web content filtering
- Advertisement blocking
- Stealth crawling
- Malicious tagging
- ... more?

# Opportunities for Web spam

## ❌ Spamdexing

- Keyword stuffing
- Link farms
- Spam blogs (splogs)
- Cloaking

## Adversarial relationship

Every undeserved gain in ranking for a spammer, is a loss of precision for the search engine.

# Naïve Web Spam

Best deal for car hire discount, LOW COST CHEAP CAR HIRE. The lowest cost self drive rental in the UK. DI \_ □

File Edit View Go Bookmarks Tools Help None ▾ My Yahoo! SK posts com Ecosofia com »

http://www.carhire.ndo.co.uk/

Tejedores del Web Spam Classification http://local...ollection=1 Best deal for car ...

**[cheap car hire call center \[details here\]](#) or complete our simple [cheap car hire enquiry form \[here\]](#) and we will call you back.**

[Cheap Auto Rental] [Cheap Airport Parking] [Cheap Travel Insurance] [Cheap Foreign Currency]  
[Cheap Flight Tickets] [Cheap Hotel Rooms] [Cheap Hostels] [Cheap Package Holidays] [Cheap Weekend Breaks]

Indexed by [Linksmatch](#)  
Terms & Conditions. Privacy Policy.  
[cheepcar.co.uk](#) copyright [cheeptravel Limited](#)©  
[cheeptravel Limited](#)© part of the DHD Group Limited

**RINGTONES, LOGOS & PICTURE MESSAGES ?  
U CAN GET THEM @ RED MONGOOSE.COM**

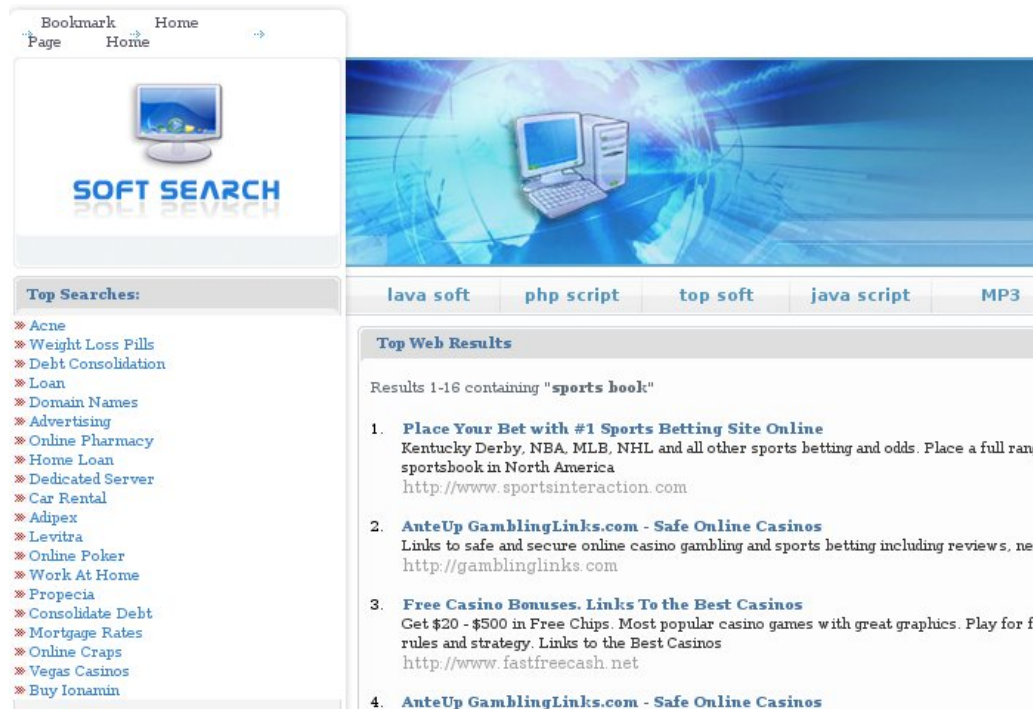
DISCOUNTED CAR HIRE IN THE UK. For the best deal on CHEAP car hire rental in the United Kingdom, visit our UK DISCOUNT SELF DRIVE feature. Guaranteed discount off normal self drive rates. DISCOUNTED CAR HIRE IN THE UK. For the best deal on CHEAP car hire rental in the United Kingdom, visit our UK DISCOUNT SELF DRIVE feature. Guaranteed discount off normal self drive rates. DISCOUNTED CAR HIRE IN THE UK. For the best deal on CHEAP car hire rental in the United Kingdom, visit our UK DISCOUNT SELF DRIVE feature. Guaranteed discount off normal self drive rates. DISCOUNTED CAR HIRE IN THE UK. For the best deal on CHEAP car hire rental in the United Kingdom, visit our UK DISCOUNT SELF DRIVE feature. Guaranteed discount off normal self drive rates. ....

Scripts Currently Forbidden [<script>: 5] [J+F+P: 0]

Done Proxy: None Adblock

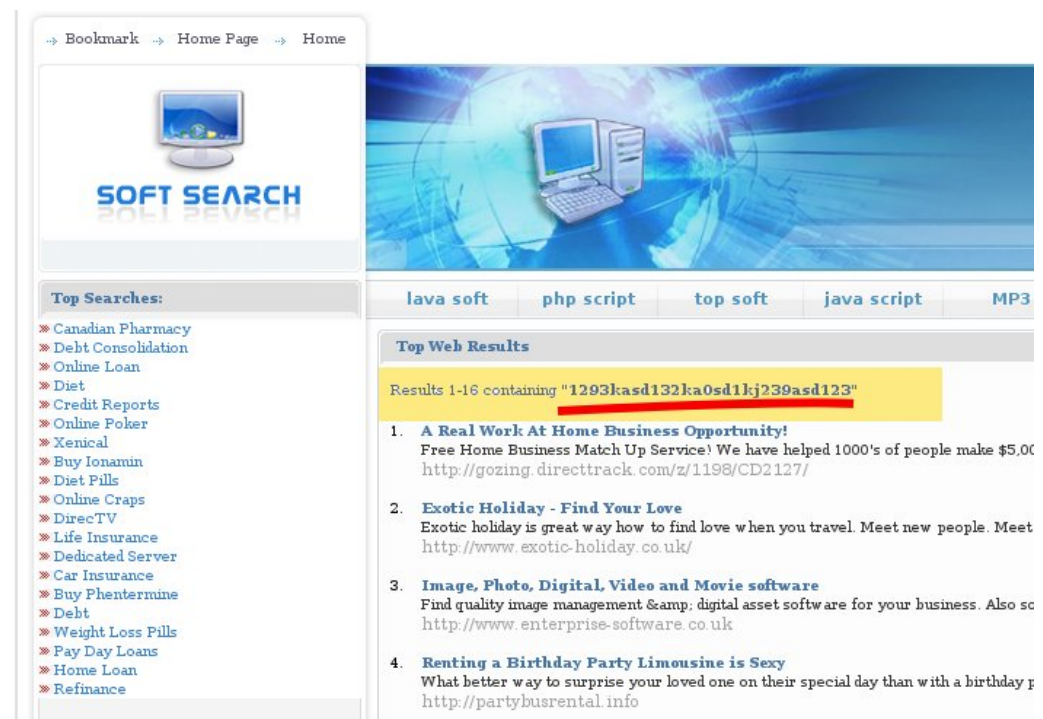


# Search engine?



The screenshot shows a search engine interface with a navigation bar at the top containing 'Bookmark', 'Home', and 'Page'. Below the navigation bar is a search bar with a computer icon and the text 'SOFT SEARCH'. To the right of the search bar is a large image of a computer monitor and keyboard. Below the search bar is a table with search results. The table has columns for 'lava soft', 'php script', 'top soft', 'java script', and 'MP3'. The search results are listed under the heading 'Top Web Results' and include links to various websites such as 'Place Your Bet with #1 Sports Betting Site Online', 'AnteUp GamblingLinks.com - Safe Online Casinos', and 'Free Casino Bonuses. Links To the Best Casinos'.

# Fake search engine



The screenshot shows a search engine interface that appears to be a fake. It has a navigation bar at the top with 'Bookmark', 'Home Page', and 'Home'. Below the navigation bar is a search bar with a computer icon and the text 'SOFT SEARCH'. To the right of the search bar is a large image of a computer monitor and keyboard. Below the search bar is a table with search results. The table has columns for 'lava soft', 'php script', 'top soft', 'java script', and 'MP3'. The search results are listed under the heading 'Top Web Results' and include links to various websites such as 'A Real Work At Home Business Opportunity!', 'Exotic Holiday - Find Your Love', 'Image, Photo, Digital, Video and Movie software', and 'Renting a Birthday Party Limousine is Sexy'. The search results are displayed in a yellow box, which is a common indicator of a fake search engine.

# “Normal” content in link farms

## 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](#)!

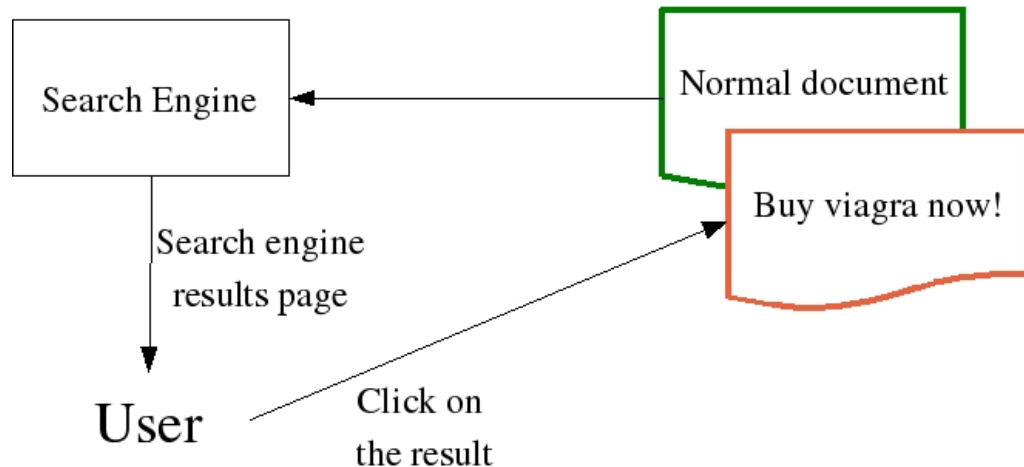
Website design, management, marketing and promotion is the key to your needs. You're one step ahead with Dry Media.

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

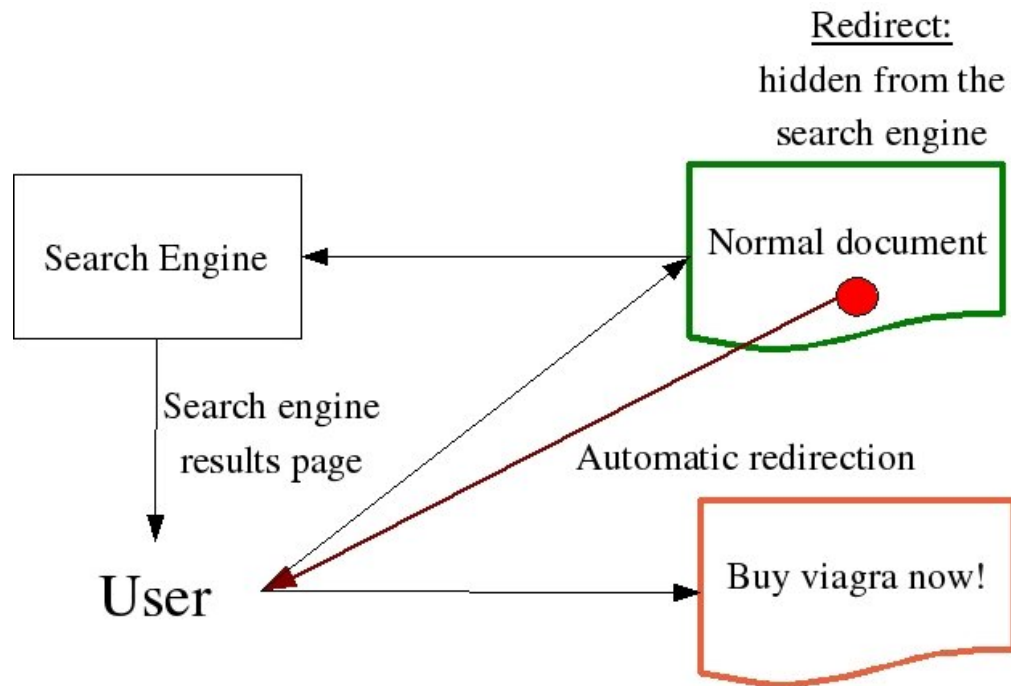
At the [Website design, management, marketing and promotion web site](#), you'll discover an easy to use, information packed source of data on Website design, management, marketing and promotion.  
[Click Here to Learn More about Website design, management, marketing and promotion.](#)

# Cloaking

Cloaking:  
different contents  
at the same URL



# Redirection



# Redirects using Javascript

## Simple redirect

```
<script>
document.location="http://www.topsearch10.com/";
</script>
```

## "Hidden" redirect

```
<script>
var1=24; var2=var1;
if(var1==var2) {
    document.location="http://www.topsearch10.com/";
}
</script>
```

## Problem: obfuscated code

### Obfuscated redirect

```
<script>
var a1="win",a2="dow",a3="loca",a4="tion.",
a5="replace",a6="('http://www.top10search.com/')";
var i,str="";
for(i=1;i<=6;i++)
{
    str += eval("a"+i);
}
eval(str);
</script>
```

## Problem: really obfuscated code

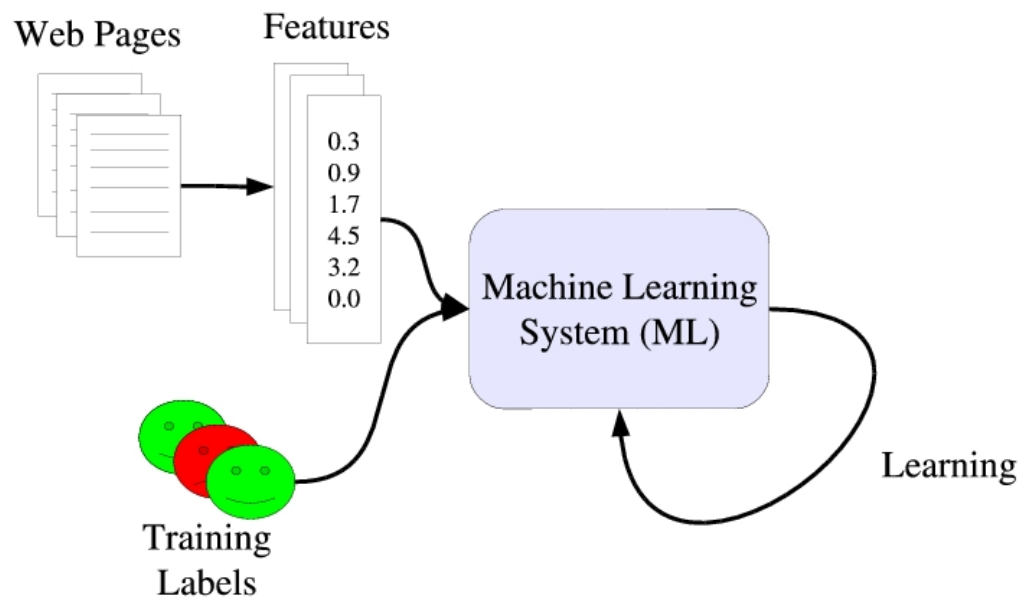
### Encoded javascript

```
<script>
var s = "%5CBE0D%5C%05GDHJ_BDE%16...%04%0E";
var e = '', i;
eval(unescape('s%eDunescape%28s%29%3Bfor...%3B')));
</script>
```

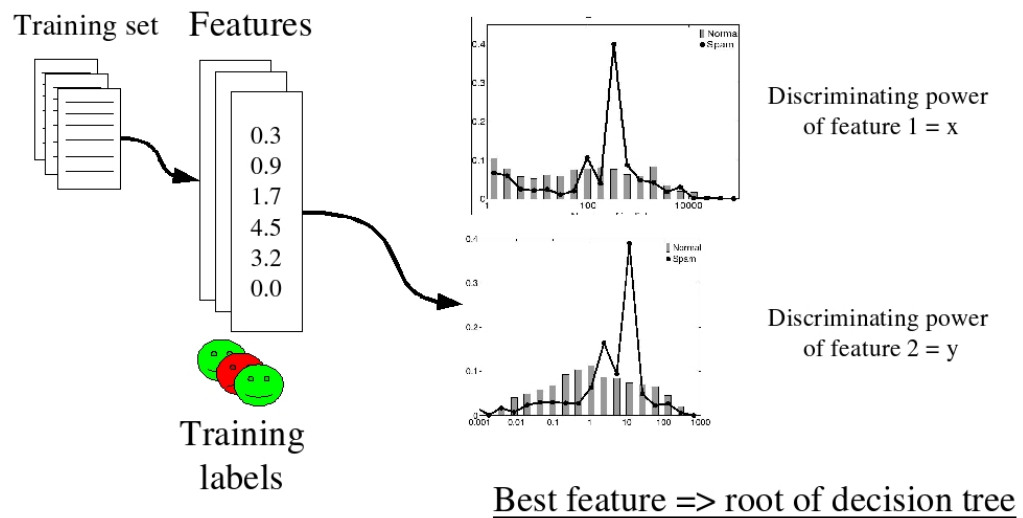
More examples: [Chellapilla and Maykov, 2007]

- 1 Web Spam
- 2 **Web Spam Detection**
- 3 A Reference Collection
- 4 Web Links
- 5 Topological Web Spam
- 6 Counting of Supporters
- 7 Content-based Spam detection
- 8 Web Topology
- 9 Conclusions

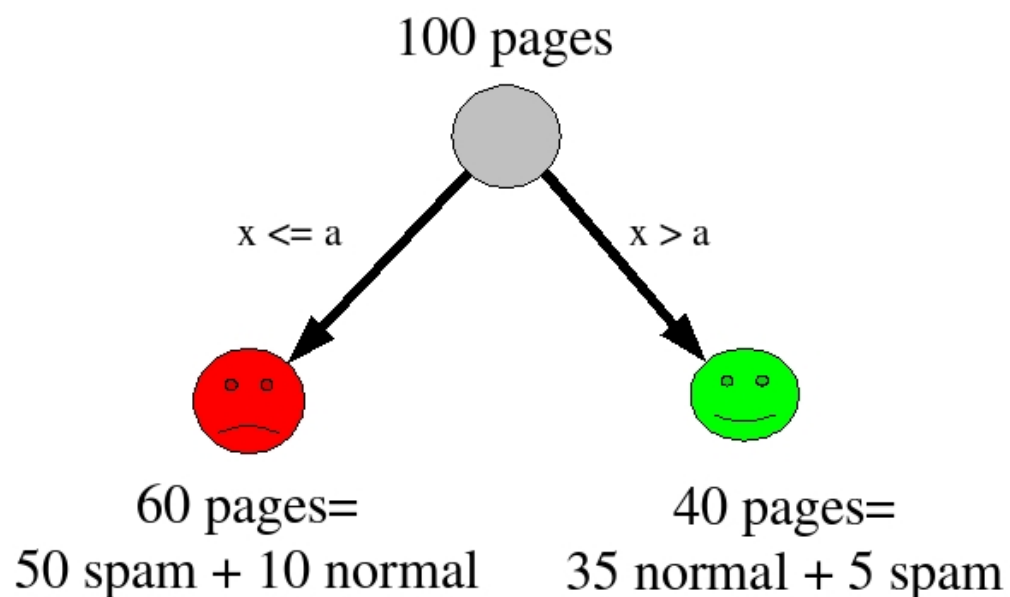
## Machine Learning



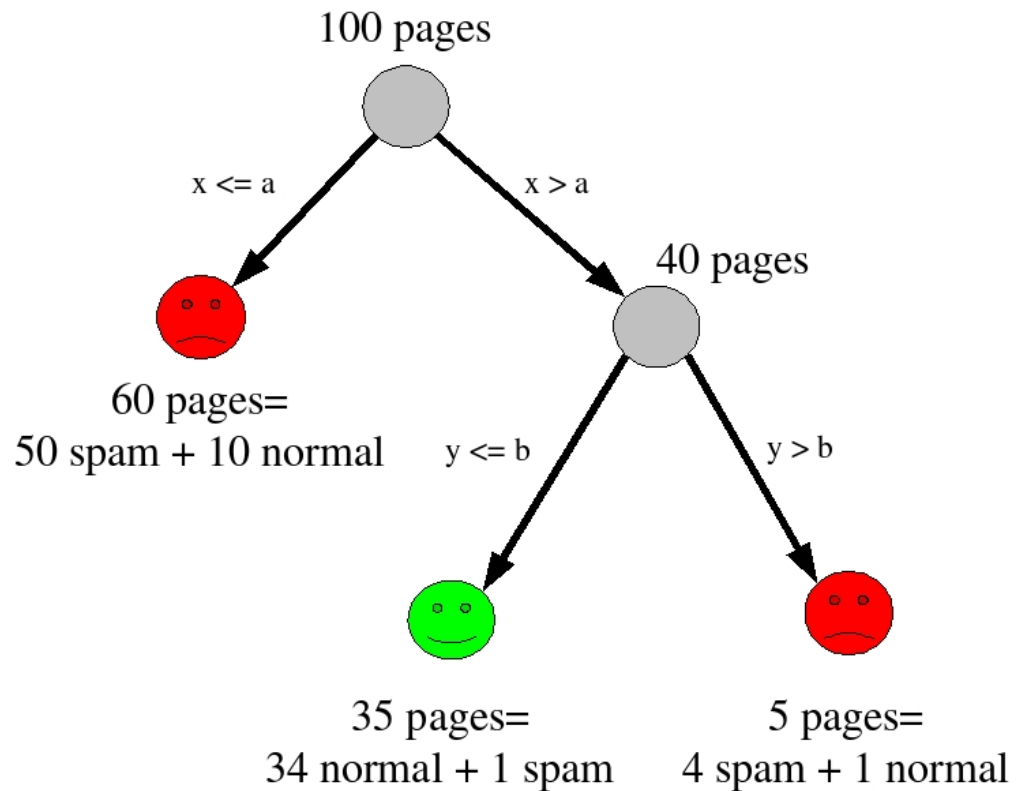
# Training of a Decision Tree



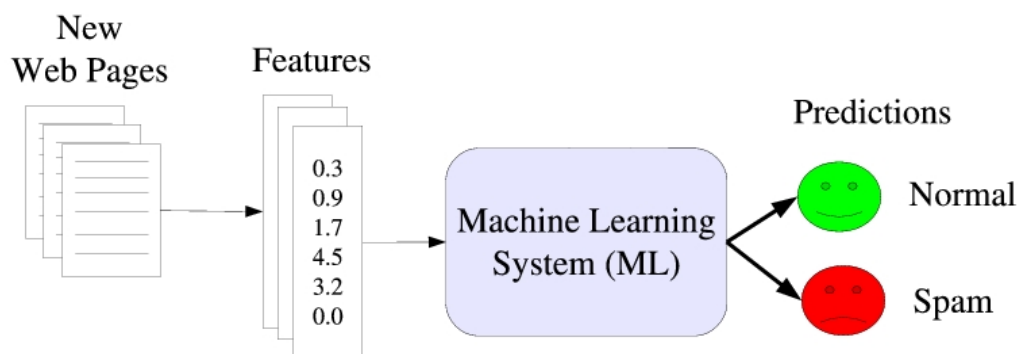
## Decision Tree (error = 15%)



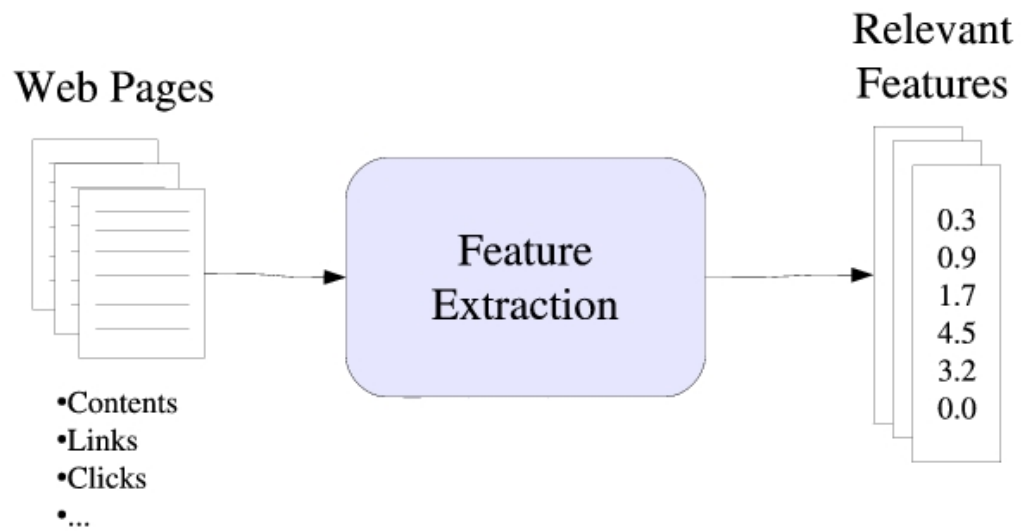
# Decision Tree (error = 15% → 12%)



# Machine Learning (cont.)



# Feature Extraction



# Challenges: Machine Learning

## Machine Learning Challenges:

- Instances are not really independent (graph)
- Learning with few examples
- Scalability

# Challenges: Information Retrieval

## Information Retrieval Challenges:

- Feature extraction: which features?
- Feature aggregation: page/host/domain
- Feature propagation (graph)
- Recall/precision tradeoffs
- Scalability

- 1 Web Spam
- 2 Web Spam Detection
- 3 A Reference Collection
- 4 Web Links
- 5 Topological Web Spam
- 6 Counting of Supporters
- 7 Content-based Spam detection
- 8 Web Topology
- 9 Conclusions

## Data is really important

- It is dangerous for a search engine to provide labelled data for this
- Even if they do, it would never reflect a consensus

## Assembling Process

- Crawling of base data
- Elaboration of the guidelines and classification interface
- Labeling
- Post-processing

# Crawling of base data

## U.K. collection

77.9 M pages downloaded from the .UK domain in May 2006  
(LAW, University of Milan)

- Large seed of about 150,000 .uk hosts
- 11,400 hosts
- 8 levels depth, with  $\leq 50,000$  pages per host

# Classification interface

The screenshot shows the 'Web Spam Test Collections - Firefox' interface. The top bar displays the URL: `http://aeserver/webspam/classify.php?workid=2#`. Below the address bar, there's a navigation pane with 'Web Spam Test Collections' selected. The main content area is divided into three panels:

- Left Panel:** A list of URLs under the heading 'Home > Collections > uk-2006-05 > Work unit'. The list includes various domains like `getme.co.uk`, `www.htm.htm.ac.uk`, `www.dazlercampani.co.uk`, `www.amsjms.mad.uk`, `ppad.ac.uk`, `www.worldhouses.co.uk`, `cookus.co.uk`, `www.select-office-services.co.uk`, `social-international.co.uk`, `www.lm-business.co.uk`, `stanuk.commission.co.uk`, `www.english.bham.ac.uk`, `visitenet.co.uk`, `dm-contracting.co.uk`, `programming.com14.ac.uk`, `www.bradford.ac.uk`, `www.hcoe.co.uk`, `cul.org.uk`, `www.directorvenues.co.uk`, and `www.cot.gov.uk`. Each entry has a status indicator (N, B, S, 7).
- Middle Panel:** A detailed view of the selected URL `www.select-office-services.co.uk`. It shows 18852 pages. The 'In-links (and PageRank contribution):' section lists several incoming links with their PageRank values. The 'Out-links (and PageRank contribution):' section lists outgoing links. The 'Extra information:' section includes a warning to use PageRank with care, a link-based PageRank of 11.45, and a traffic ranking of 4/10 from Alexa.
- Right Panel:** A preview of the webpage content for `www.select-office-services.co.uk`. It features a large orange 'office broker .com' logo, a phone number '0800 111 6', and the text 'the best place'. Below this, there's a 'Welcome Off' section and a search bar for offices in the UK.

At the bottom of the interface, there are buttons for 'Save Scores', 'Done', 'Cancel - I will work on this later', and 'Delete - I won't do this work unit'.

# Labeling process

- We asked 20+ volunteers to classify entire hosts
- Asked to classify **normal** / **borderline** / **spam**
- Do they agree? Mostly ...

# Agreement

2547	<a href="#">infomove.gub.ar.uk</a>	AUTO_damian N	
2548	<a href="#">info.hut.ac.uk</a>	AUTO_damian N	AUTO_damian N
2549	<a href="#">infoportcity/notes.dahermscuprima.co.uk</a>	joanne S	joanne S
2550	<a href="#">info@adnet.co.uk</a>	AUTO_damian N	searung N
2551	<a href="#">info@nrmrta.dnrc.co.uk</a>	antonio N	chato B
2552	<a href="#">insuranceonlineukdays.org.uk</a>	xiexiang N	jamie S
2553	<a href="#">inreach.ac.uk</a>	AUTO_damian N	AUTO_damian N
2554	<a href="#">inreach.brighton.ac.uk</a>	AUTO_damian N	
2555	<a href="#">inreach.bath.ac.uk</a>	AUTO_damian N	
2556	<a href="#">inreach.co.net.ac.uk</a>	AUTO_damian N	
2557	<a href="#">inreach.sps.kel.ac.uk</a>	AUTO_damian N	
2558	<a href="#">inreachinspirationes.co.uk</a>	thomas S	antonio N
2559	<a href="#">inrent-nfm.ac.uk</a>	AUTO_damian N	AUTO_damian N
2560	<a href="#">inrent.bham.ac.uk:259</a>	antonio ?	chato N
2561	<a href="#">inrent.wa.ac.uk</a>	AUTO_damian N	
2562	<a href="#">inrent.london.ac.uk</a>	AUTO_damian N	
2563	<a href="#">inrent.open.ac.uk</a>	AUTO_damian N	
2564	<a href="#">inrent.sufford.ac.uk</a>	AUTO_damian N	
2565	<a href="#">investing.madison.co.uk</a>	thomas N	mike N
2566	<a href="#">investing.theirmoney.co.uk</a>	min N	alex B
2567	<a href="#">investorcenter.lsa.ac.uk</a>	AUTO_damian N	AUTO_damian N
2568	<a href="#">i2i.ac.uk</a>	omar N	zlatan N
2569	<a href="#">i2i.ac.uk</a>	AUTO_damian N	AUTO_damian N
2570	<a href="#">i2i.ac.uk</a>	AUTO_damian N	
2571	<a href="#">i2i.ac.uk</a>	AUTO_damian N	
2572	<a href="#">i2i.ac.uk</a>	min S	lucy ?
2573	<a href="#">i2i.ac.uk</a>	piecel N	zlatan N
2574	<a href="#">i2i.ac.uk</a>	AUTO_damian N	AUTO_damian N
2575	<a href="#">i2i.ac.uk</a>	larry S	zlatan S
2576	<a href="#">i2i.ac.uk</a>	AUTO_damian N	
2577	<a href="#">i2i.ac.uk</a>	antonio S	antonio S
2578	<a href="#">i2i.ac.uk</a>	maximo N	ben N
2579	<a href="#">i2i.ac.uk</a>	xiexiang N	thego N

# Results

## Labels

Label	Frequency	Percentage
Normal	4,046	61.75%
Borderline	709	10.82%
Spam	1,447	22.08%
Can not classify	350	5.34%

## Agreement

Category	Kappa	Interpretation
normal	0.62	Substantial agreement
spam	0.63	Substantial agreement
borderline	0.11	Slight agreement
global	0.56	Moderate agreement

# Result: first public Web Spam collection

- Public spam collection
  - Labels for 6,552 hosts
  - 2,725 hosts classified by at least 2 humans
  - 3,106 automatically considered normal (.ac.uk, .sch.uk, .gov.uk, .mod.uk, .nhs.uk or .police.uk)
  - <http://www.yr-bcn.es/webspam/>
- Upcoming Web Spam challenge
  - Track I: Information retrieval + Machine learning
  - Track II: Machine learning
  - <http://webspam.lip6.fr/>
- AIRWeb 2007 Workshop (challenge results available)
  - Regular and short papers
  - Track I of the Web Spam Challenge
  - <http://airweb.cse.lehigh.edu/2007/>

# AIRWeb 2007 in Banff, Canada

The screenshot shows the homepage of the AIRWeb 2007 workshop. At the top, there is a navigation bar with links: Home, Call for Papers, Submissions, Program, and Contact. To the right of the navigation bar, it says "May 8th, 2007 — Banff, Alberta, CANADA." Below the navigation bar is a large banner image of a snow-capped mountain. Overlaid on the banner is the text "AIRWeb 2007" and "Third International Workshop on Adversarial Information Retrieval on the Web". Below the banner, the main content area has a "Welcome!" section. It describes AIRWeb as a series of international workshops focusing on Adversarial Information Retrieval on the Web, bringing together researchers and industry practitioners. It mentions that the 2007 workshop is co-located with the WWW07 conference in Banff, Canada, and will include a Web Spam challenge. To the right of the main text, there is a sidebar with a section titled "AIRWEB'07" containing links: Home, Call for Papers, Submissions, Program, and Contact. Below that is a section titled "PAST WORKSHOPS" listing AIRWeb'06 (Seattle, USA), AIRWeb'05 (Chiba, Japan), and Chiba, Japan.

Home Call for Papers Submissions Program Contact May 8th, 2007 — Banff, Alberta, CANADA.

## AIRWeb 2007

### Third International Workshop on Adversarial Information Retrieval on the Web

## Welcome!

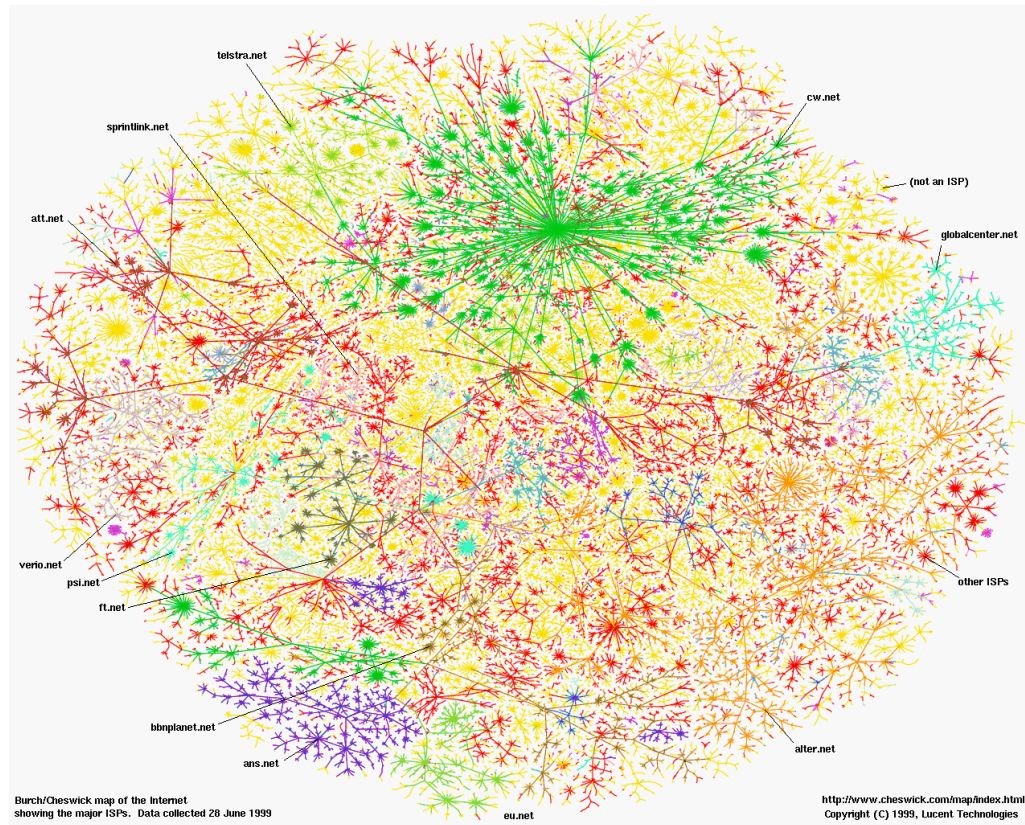
AIRWeb is a series of international workshops focusing on Adversarial Information Retrieval on the Web that brings together both researchers and industry practitioners, to present and discuss advances in the state of the art.

This year, AIRWeb2007 will be co-located with the WWW07 conference in Banff, Canada. The workshop will include a Web Spam challenge that will test different spam detection techniques on a shared reference collection. Accepted papers will be posted on the ACM Digital Library.

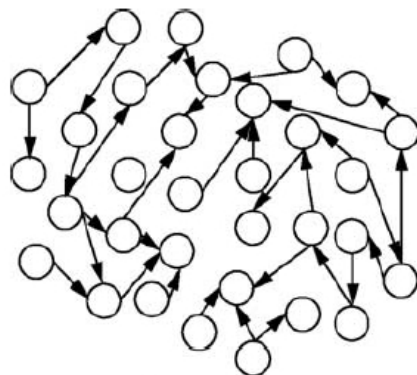
**AIRWEB'07**  
Home  
Call for Papers  
Submissions  
Program  
Contact

**PAST WORKSHOPS**  
AIRWeb'06  
Seattle, USA  
AIRWeb'05  
Chiba, Japan

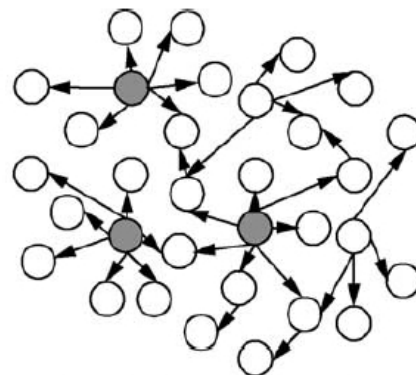
- 1 Web Spam
- 2 Web Spam Detection
- 3 A Reference Collection
- 4 Web Links
- 5 Topological Web Spam
- 6 Counting of Supporters
- 7 Content-based Spam detection
- 8 Web Topology
- 9 Conclusions



## Scale-free networks



(a) Random network



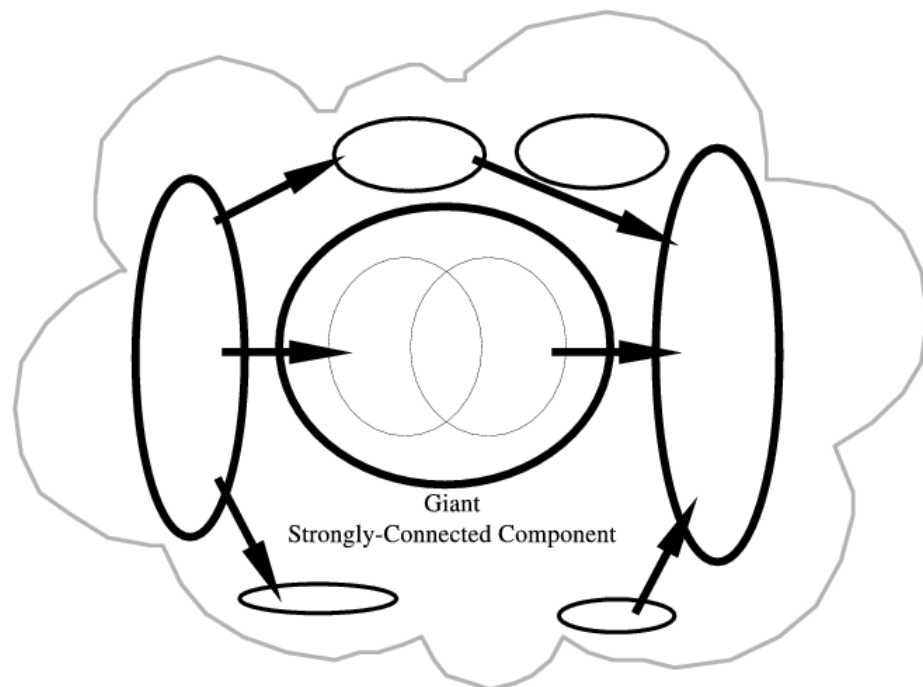
(b) Scale-free network

# How to find meaningful patterns?

Several levels of analysis:

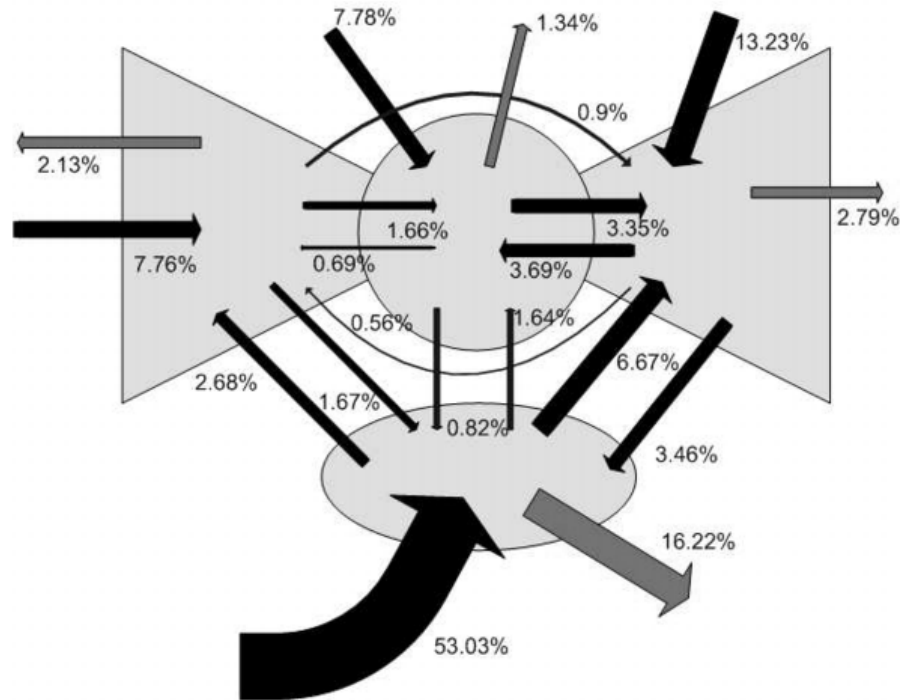
- Macroscopic view: overall structure
- Microscopic view: nodes
- Mesoscopic view: regions

## Macroscopic view, e.g. Bow-tie



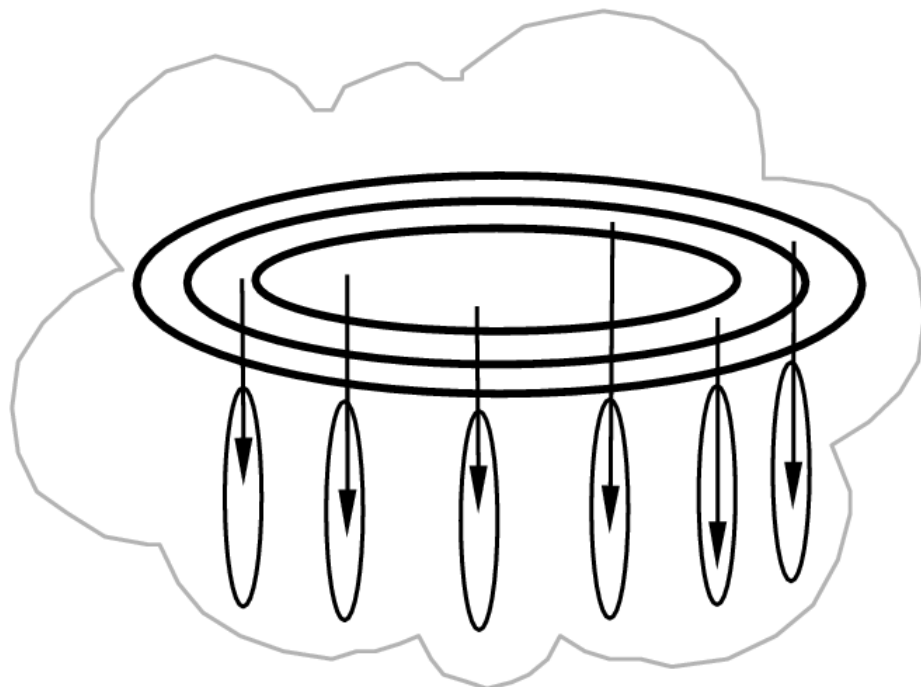
[Broder et al., 2000]

## Macroscopic view, e.g. Bow-tie, migration



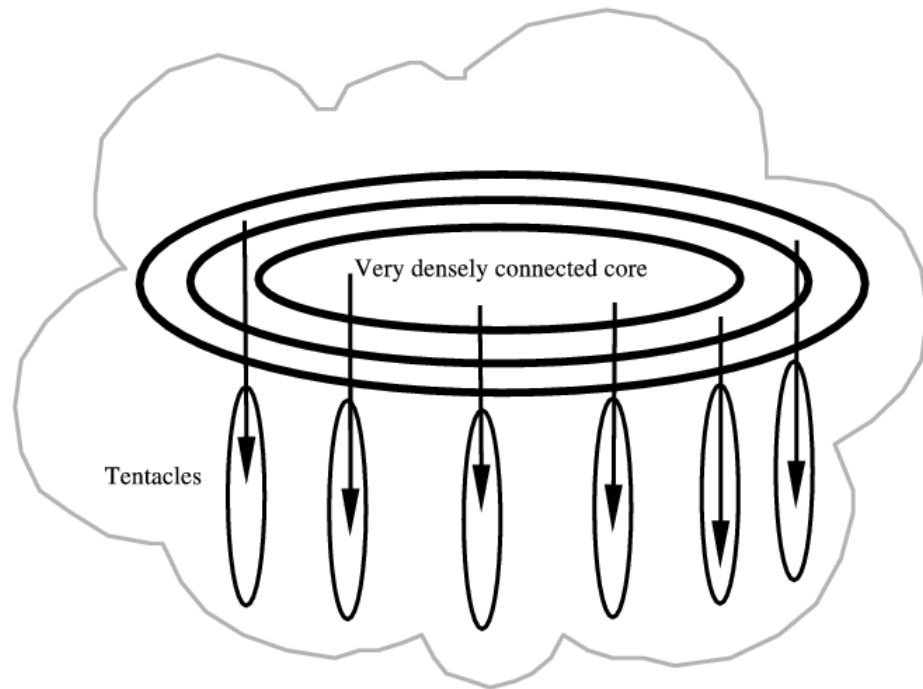
[Baeza-Yates and Poblete, 2006]

## Macroscopic view, e.g. Jellyfish

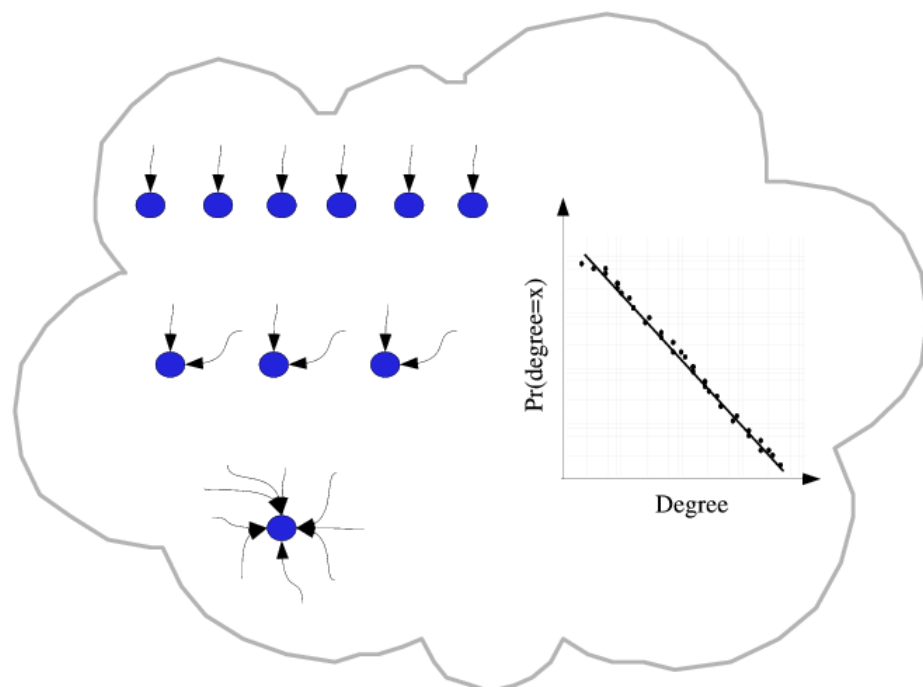


[Tauro et al., 2001] - Internet Autonomous Systems (AS)  
Topology

## Macroscopic view, e.g. Jellyfish



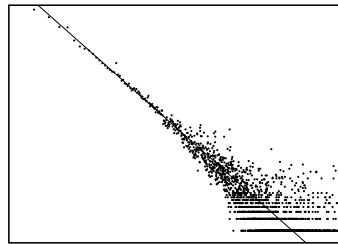
## Microscopic view, e.g. Degree



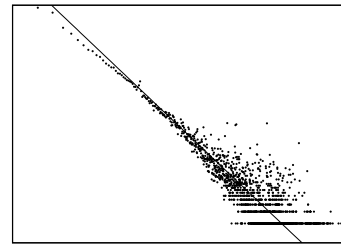
[Barabási, 2002] and others

## Microscopic view, e.g. Degree

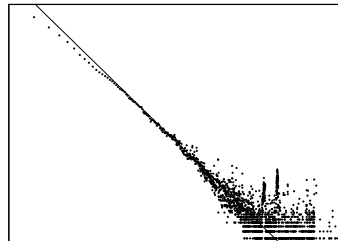
Greece



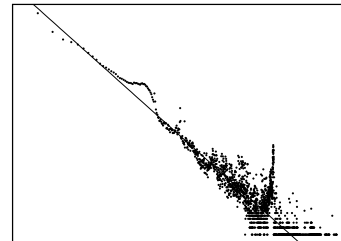
Chile



Spain

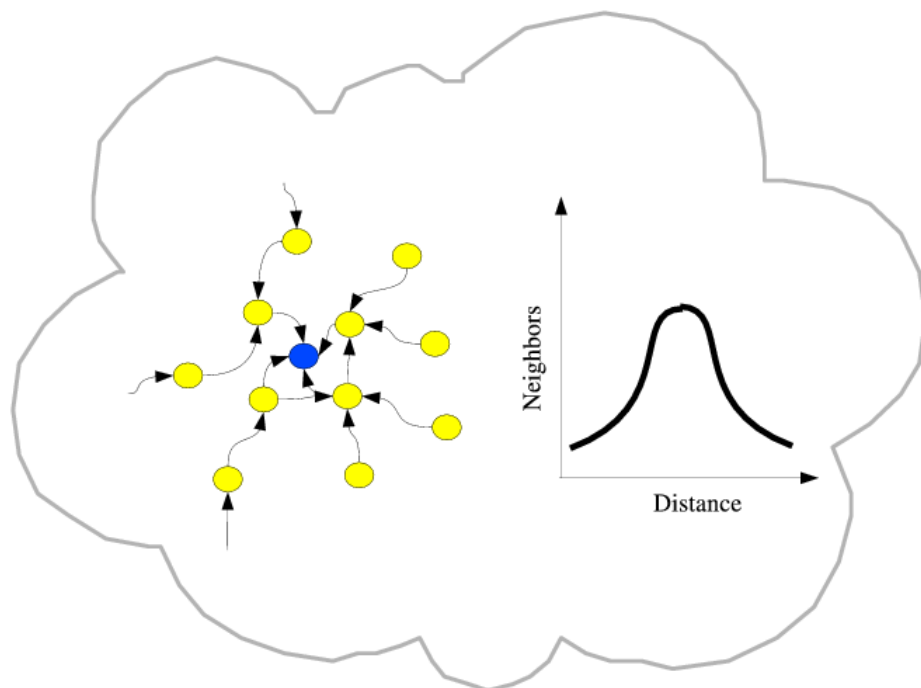


Korea

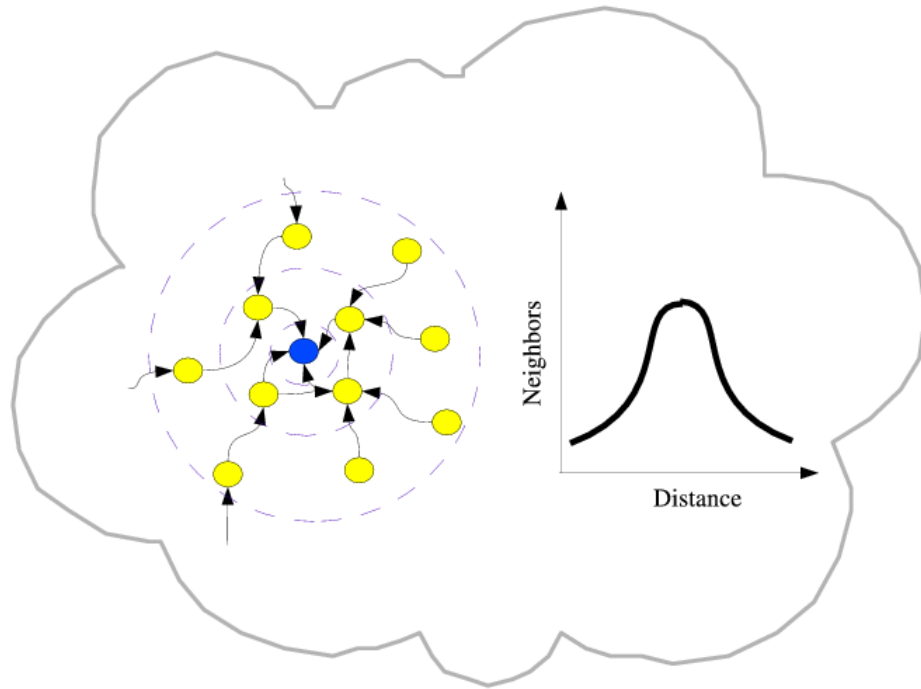


[Baeza-Yates et al., 2006b] - compares this distribution in 8 countries ... guess what is the result?

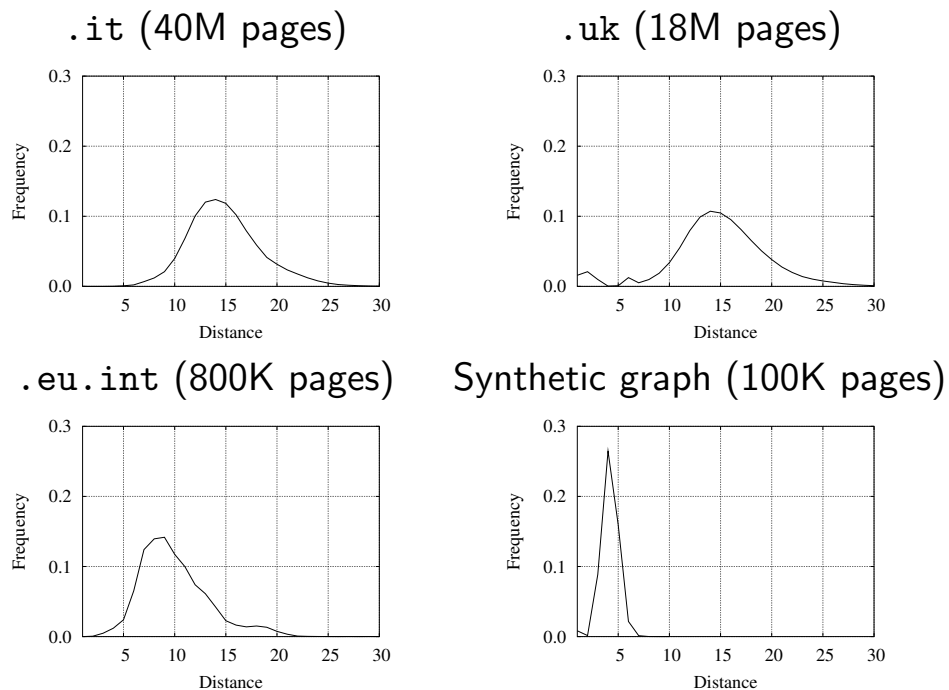
## Mesoscopic view, e.g. Hop-plot



## Mesoscopic view, e.g. Hop-plot

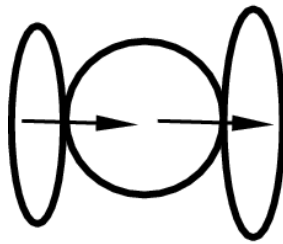


## Mesoscopic view, e.g. Hop-plot



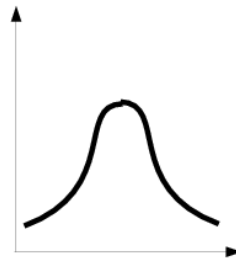
[Baeza-Yates et al., 2006a]

## Macroscopic



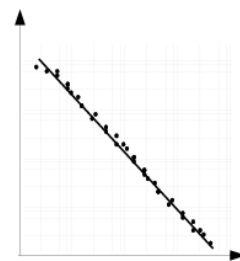
Connected components  
Jellyfish structure  
Bow-tie structure  
...

## Mesoscopic



Hop-plots  
Link-based ranking  
Clusters, communities  
...

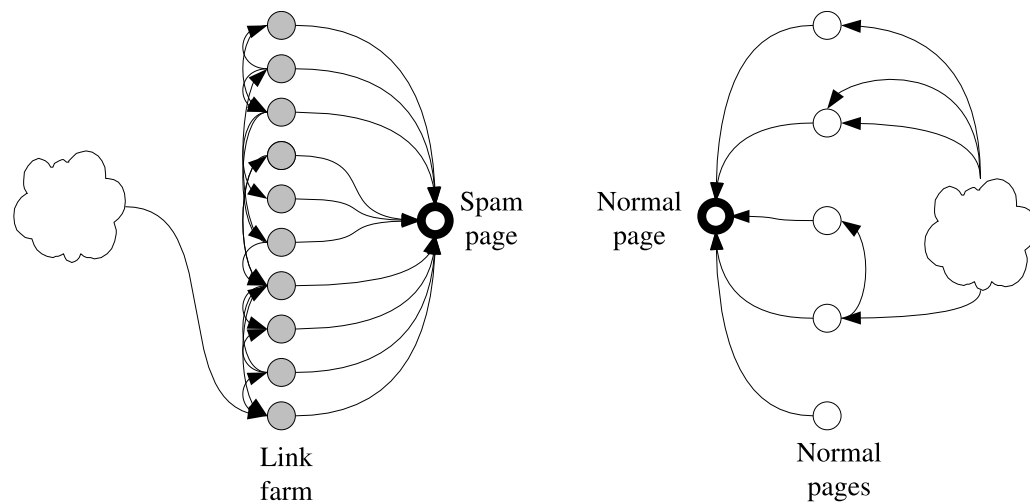
## Microscopic



Zipf's law  
Degree distributions  
...

- 1 Web Spam
- 2 Web Spam Detection
- 3 A Reference Collection
- 4 Web Links
- 5 Topological Web Spam
- 6 Counting of Supporters
- 7 Content-based Spam detection
- 8 Web Topology
- 9 Conclusions

# Topological spam: link farms



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

## Motivation

Fetterly [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”**

# Handling large graphs

For large graphs, random access is not possible.

Large graphs do not fit in main memory

Streaming model of computation

# Semi-streaming model

- Memory size enough to hold some data per-node
- Disk size enough to hold some data per-edge
- A small number of passes over the data

# Restriction

**Semi-streaming model:** graph on disk

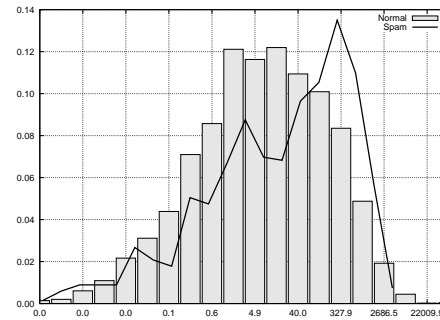
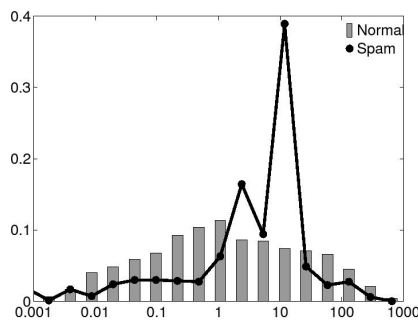
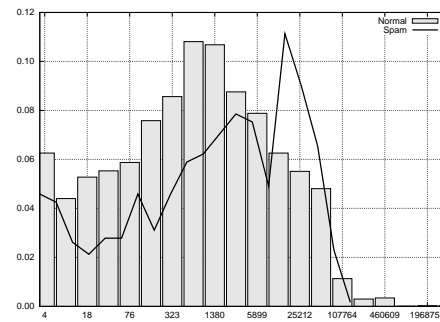
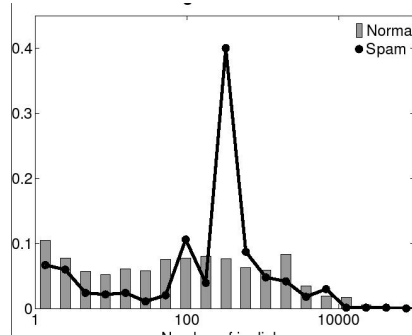
```
1: for node : 1 ... N do
2:   INITIALIZE-MEM(node)
3: end for
4: for distance : 1 ... d do {Iteration step}
5:   for src : 1 ... N do {Follow links in the graph}
6:     for all links from src to dest do
7:       COMPUTE(src,dest)
8:     end for
9:   end for
10:  NORMALIZE
11: end for
12: POST-PROCESS
13: return Something
```

# Link-Based Features

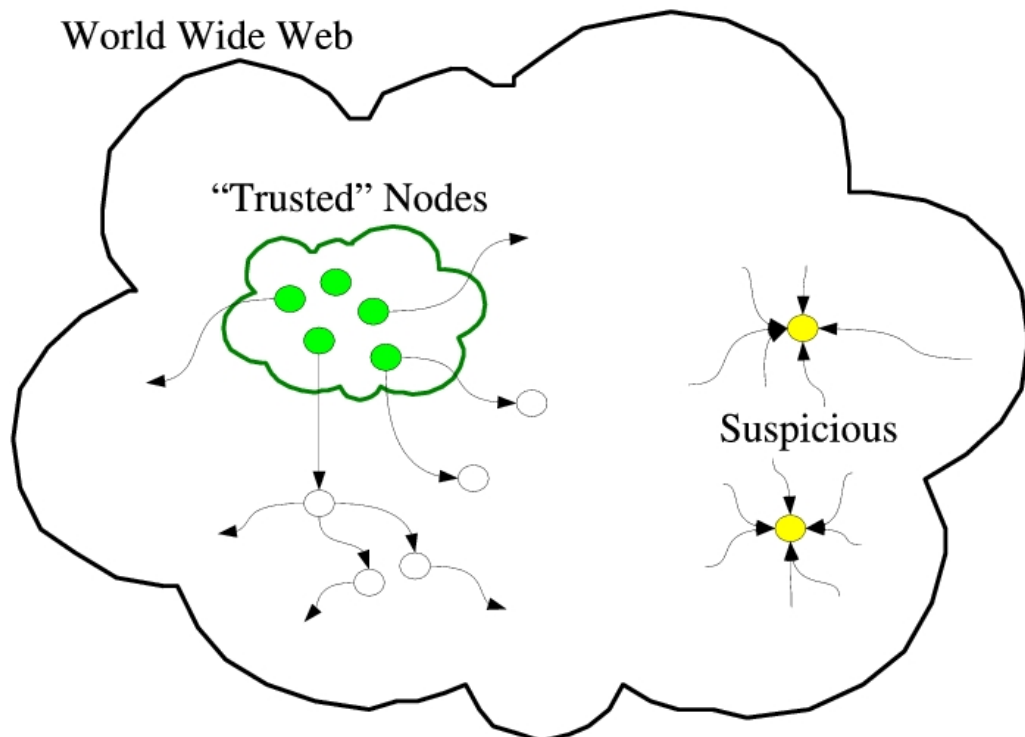
- Degree-related measures
- PageRank
- TrustRank [Gyöngyi et al., 2004]
- Truncated PageRank [Becchetti et al., 2006]
- Estimation of supporters [Becchetti et al., 2006]

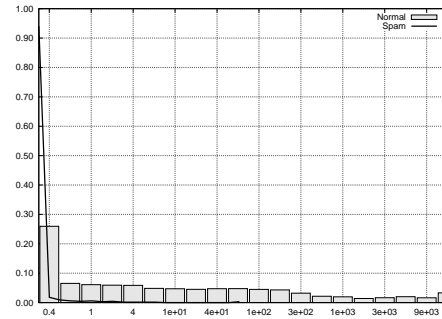
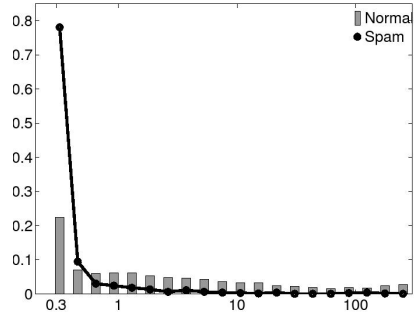
140 features per host (2 pages per host)

## Degree-Based



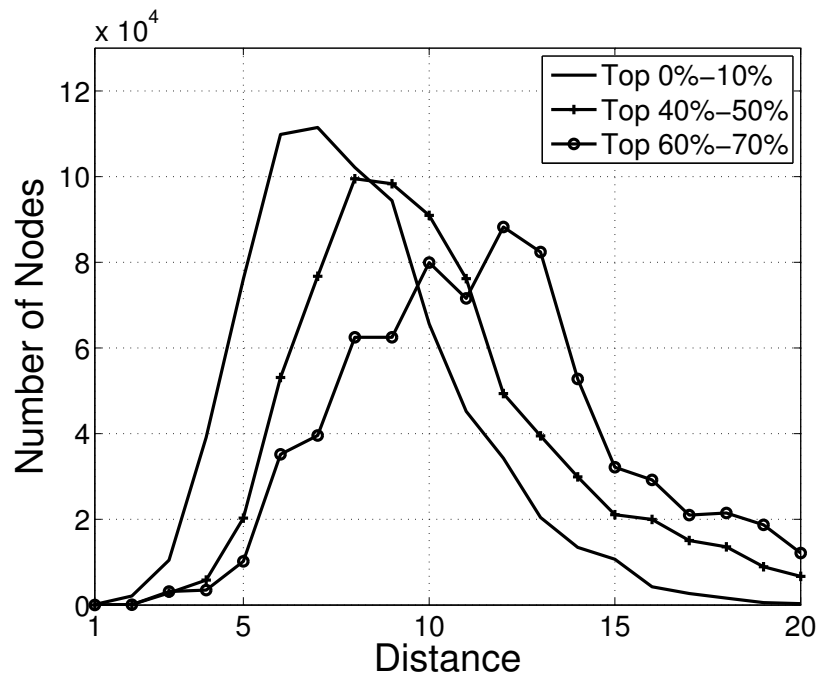
## TrustRank Idea





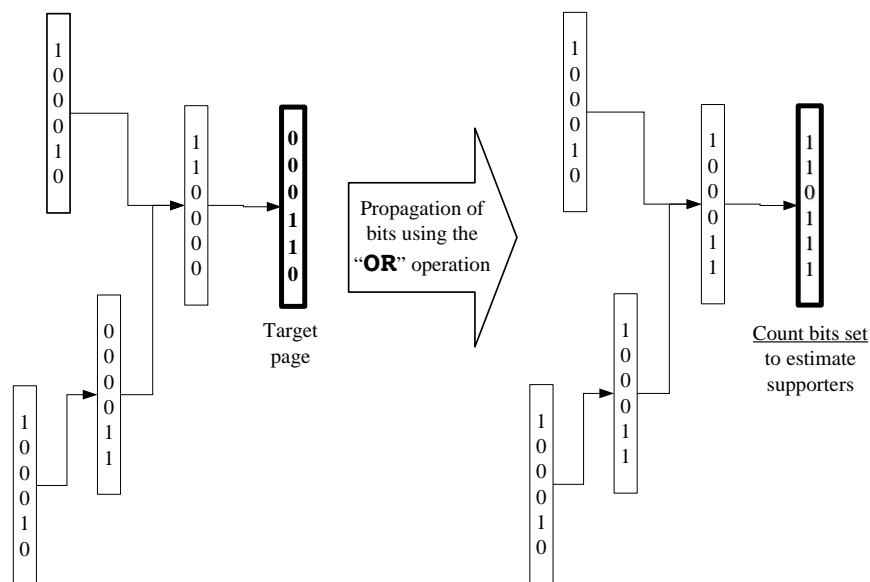
- 1 Web Spam
- 2 Web Spam Detection
- 3 A Reference Collection
- 4 Web Links
- 5 Topological Web Spam
- 6 Counting of Supporters
- 7 Content-based Spam detection
- 8 Web Topology
- 9 Conclusions

# High and low-ranked pages are different



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

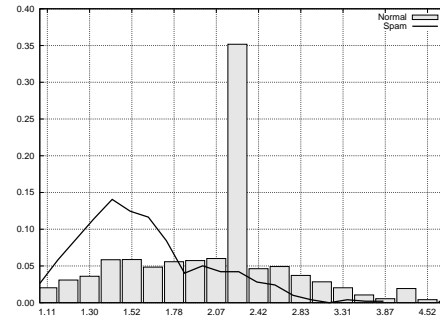
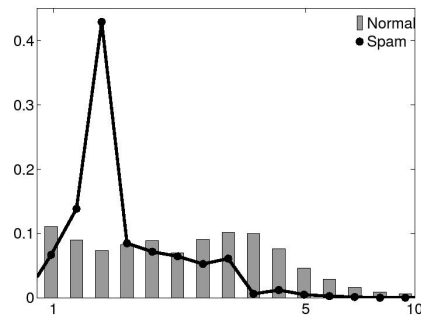
# Probabilistic counting



[Becchetti et al., 2006] shows an improvement of ANF algorithm [Palmer et al., 2002] based on probabilistic counting [Flajolet and Martin, 1985]

# Bottleneck number

$b_d(x) = \min_{j \leq d} \{|N_j(x)| / |N_{j-1}(x)|\}$ . Minimum rate of growth of the neighbors of  $x$  up to a certain distance. We expect that spam pages form clusters that are somehow isolated from the rest of the Web graph and they have smaller bottleneck numbers than non-spam pages.



- 1 Web Spam
- 2 Web Spam Detection
- 3 A Reference Collection
- 4 Web Links
- 5 Topological Web Spam
- 6 Counting of Supporters
- 7 Content-based Spam detection
- 8 Web Topology
- 9 Conclusions

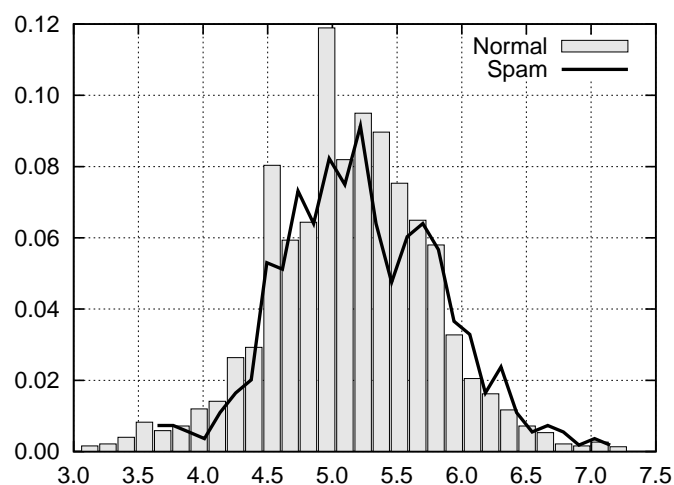
# Content-Based Features

Most of the features reported in [Ntoulas et al., 2006]

- Number of word in the page and title
- Average word length
- Fraction of anchor text
- Fraction of visible text
- Compression rate
- Corpus precision and corpus recall
- Query precision and query recall
- Independent trigram likelihood
- Entropy of trigrams

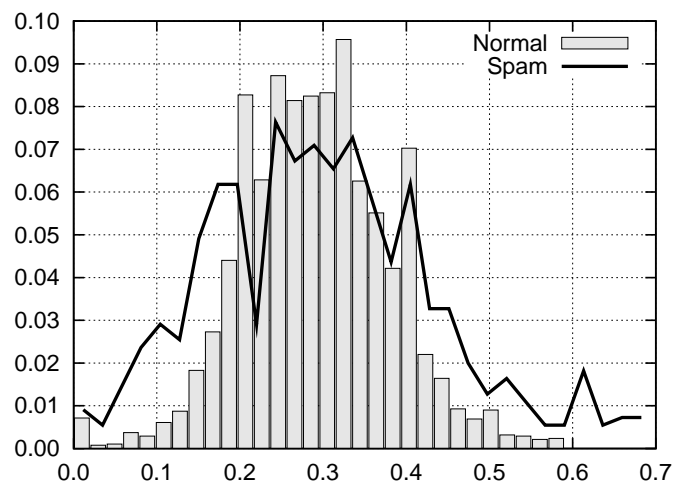
96 features per host

## Average word length



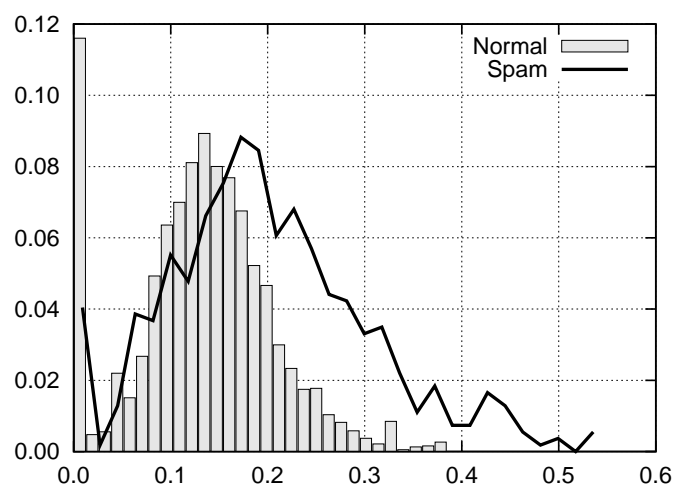
**Figure:** Histogram of the average word length in non-spam vs. spam pages for  $k = 500$ .

# Corpus precision



**Figure:** Histogram of the corpus precision in non-spam vs. spam pages.

# Query precision



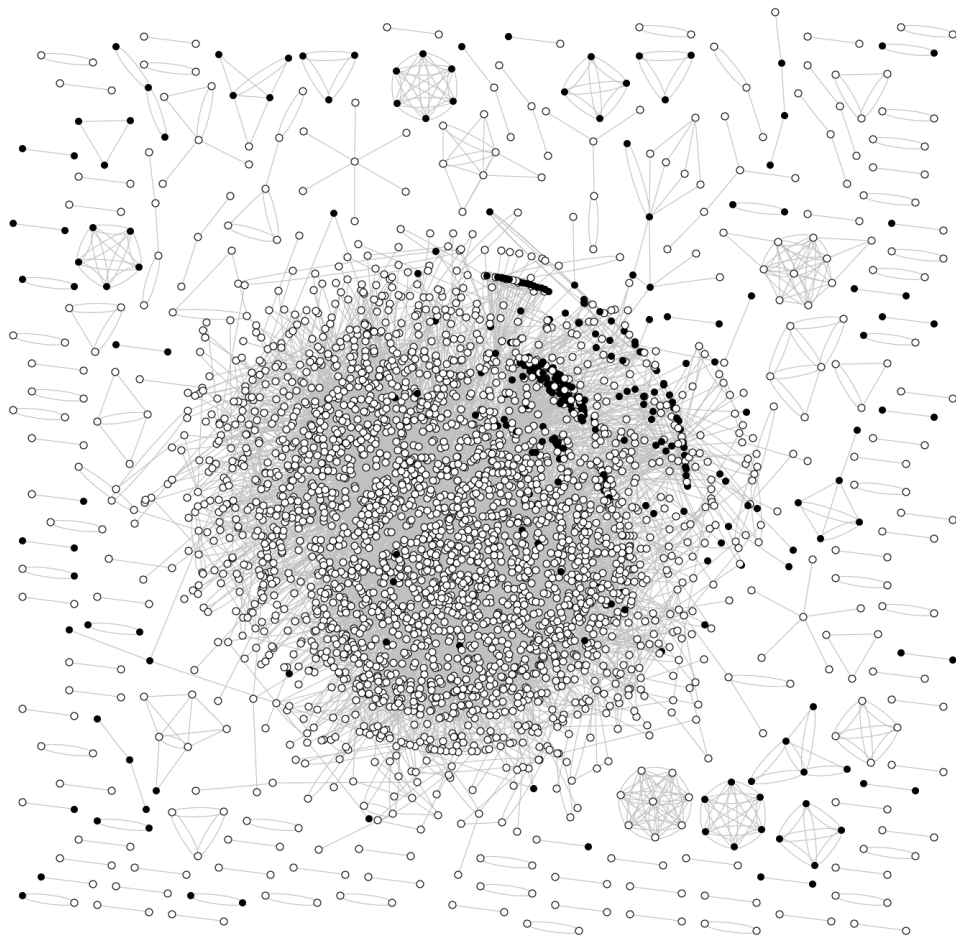
**Figure:** Histogram of the query precision in non-spam vs. spam pages for  $k = 500$ .

- 1 Web Spam
- 2 Web Spam Detection
- 3 A Reference Collection
- 4 Web Links
- 5 Topological Web Spam
- 6 Counting of Supporters
- 7 Content-based Spam detection
- 8 **Web Topology**
- 9 Conclusions

## General hypothesis

**Pages topologically close to each other are more likely to have the same label (spam/nospam) than random pairs of pages.**

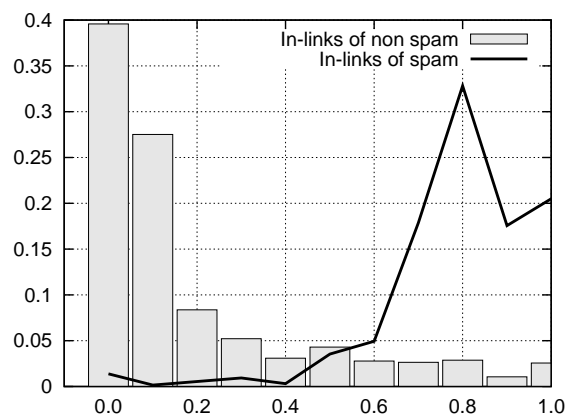
Pages linked together are more likely to be on the same topic than random pairs of pages [Davison, 2000]



## Topological dependencies: in-links

Histogram of fraction of spam hosts in the in-links

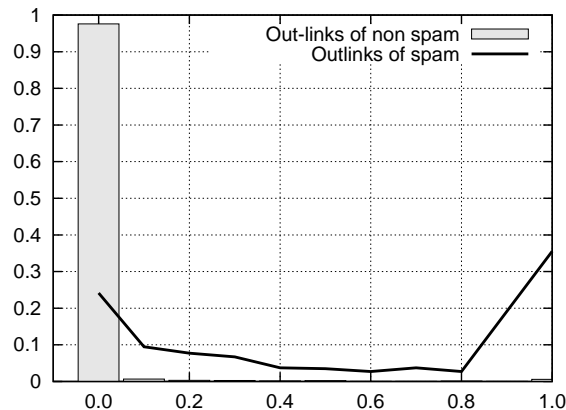
- 0 = no in-link comes from spam hosts
- 1 = all of the in-links come from spam hosts



# Topological dependencies: out-links

Histogram of fraction of spam hosts in the out-links

- 0 = none of the out-links points to spam hosts
- 1 = all of the out-links point to spam hosts



## Idea 1: Clustering

Classify, then cluster hosts, then assign the same label to all hosts in the same cluster by majority voting

	Baseline	Clustering
Without bagging		
True positive rate	75.6%	74.5%
False positive rate	8.5%	6.8%
F-Measure	0.646	<b>0.673</b>
With bagging		
True positive rate	78.7%	76.9%
False positive rate	5.7%	5.0%
F-Measure	0.723	0.728

✓ Reduces error rate

## Idea 2: Propagate the label

Classify, then interpret “spamcity” as a probability, then do a random walk with restart from those nodes

	Baseline	Fwds.	Backwds.	Both
Classifier without bagging				
True positive rate	75.6%	70.9%	69.4%	71.4%
False positive rate	8.5%	6.1%	5.8%	5.8%
F-Measure	0.646	0.665	0.664	<b>0.676</b>
Classifier with bagging				
True positive rate	78.7%	76.5%	75.0%	75.2%
False positive rate	5.7%	5.4%	4.3%	4.7%
F-Measure	0.723	0.716	0.733	0.724

## Idea 3: Stacked graphical learning

Classify, then add the average predicted “spamcity” of neighbors as a new feature for each node, then classify again[Cohen and Kou, 2006]

	Baseline	Avg. of in	Avg. of out	Avg. of both
True positive rate	78.7%	84.4%	78.3%	85.2%
False positive rate	5.7%	6.7%	4.8%	6.1%
F-Measure	0.723	0.733	0.742	<b>0.750</b>

✓ Increases detection rate

# Idea 3: Stacked graphical learning x2

And repeat ...

	Baseline	First pass	Second pass
True positive rate	78.7%	85.2%	88.4%
False positive rate	5.7%	6.1%	6.3%
F-Measure	0.723	0.750	<b>0.763</b>

✓ Significant improvement over the baseline

- 1 Web Spam
- 2 Web Spam Detection
- 3 A Reference Collection
- 4 Web Links
- 5 Topological Web Spam
- 6 Counting of Supporters
- 7 Content-based Spam detection
- 8 Web Topology
- 9 Conclusions

## Concluding remarks

- ✓ The UK-2006-05 dataset is “harder” than previous datasets
- ✓ Considering content-based and link-based attributes improves the accuracy
- ✓ Considering the dependencies improves the accuracy

Thank you!

## Web Spam Detection

R. Baeza-Yates

Web Spam

Web Spam Detection

A Reference Collection

Web Links

Topological Web Spam

Counting of Supporters

Content-based Spam detection

Web Topology

Conclusions



Baeza-Yates, R., Boldi, P., and Castillo, C. (2006a).  
Generalizing pagerank: Damping functions for link-based ranking algorithms.  
*In Proceedings of ACM SIGIR*, pages 308–315, Seattle, Washington, USA. ACM Press.



Baeza-Yates, R., Castillo, C., and Efthimiadis, E. (2006b).  
Characterization of national web domains.  
*To appear in ACM TOIT*.



Baeza-Yates, R. and Poblete, B. (2006).  
Dynamics of the chilean web structure.  
*Comput. Networks*, 50(10):1464–1473.



Barabási, A.-L. (2002).  
*Linked: The New Science of Networks*.  
Perseus Books Group.



Becchetti, L., Castillo, C., Donato, D., Leonardi, S., and Baeza-Yates, R. (2006).  
Using rank propagation and probabilistic counting for link-based spam detection.  
*In Proceedings of the Workshop on Web Mining and Web Usage Analysis (WebKDD)*, Pennsylvania, USA. ACM Press.

## Web Spam Detection

R. Baeza-Yates

Web Spam

Web Spam Detection

A Reference Collection

Web Links

Topological Web Spam

Counting of Supporters

Content-based Spam detection

Web Topology

Conclusions



Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000).  
Graph structure in the web: Experiments and models.  
*In Proceedings of the Ninth Conference on World Wide Web*, pages 309–320, Amsterdam, Netherlands. ACM Press.



Chellapilla, K. and Maykov, A. (2007).  
A taxonomy of javascript redirection spam.  
*In AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 81–88, New York, NY, USA. ACM Press.



Cohen, W. W. and Kou, Z. (2006).  
Stacked graphical learning: approximating learning in markov random fields using very short inhomogeneous markov chains.  
Technical report.



Davison, B. D. (2000).  
Topical locality in the web.  
*In Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*, pages 272–279, Athens, Greece. ACM Press.



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.



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

Combating Web spam with TrustRank.

In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, pages 576–587, Toronto, Canada. Morgan Kaufmann.



Ntoulas, A., Najork, M., Manasse, M., and Fetterly, D. (2006).

Detecting spam web pages through content analysis.

In *Proceedings of the World Wide Web conference*, pages 83–92, 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.



Tauro, L., Palmer, C., Siganos, G., and Faloutsos, M. (2001).

A simple conceptual model for the internet topology.

In *Global Internet*, San Antonio, Texas, USA. IEEE CS Press.