Web Spam

Know Your Neighbors: Web Spam Detection using the Web Topology

Presenter: Sadia Masood

Tutor : Klaus Berberich

Date : 17-Jan-2008

The Agenda

- Focus of the First Paper
- Motivation
- The Objective
- DATASET
- Link Based Features
- Content Based Features
- Using the Content & Link Based Features
- Using the Web Topology
- Conclusion

The Agenda

- Focus of the First Paper
- **Motivation**
- The Objective
- DATASET
- Link Based Features
- Content Based Features
- Using the Content & Link Based Features
- Using the Web Topology
- Conclusion

On The Web

Includes Information + Free Movies + Hotel Reservation + Casino + Buy Cheap Software + Flight Bookings + Win Lottery & etc

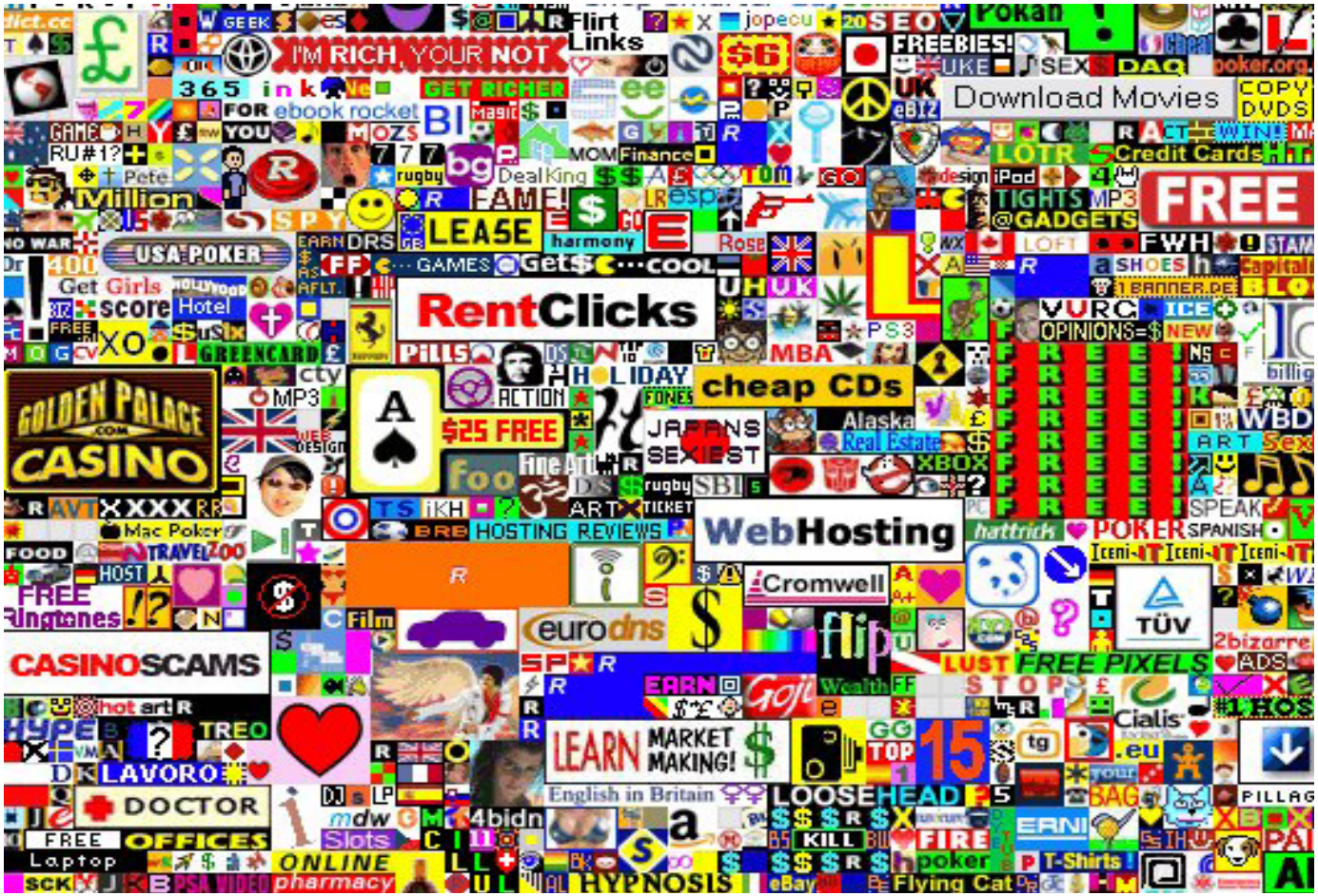


Image: www.milliondollarhomepage.com

What is Web Spamming?

"**Any deliberate human action that is meant to trigger an unjustifiably favorable relevance or importance of some web page considering the page's true value**" [5]

Also called ***Search Engine Spamming*** or ***Spamdexing***

[5]: Z. Gyongyi and H. Garcia-Molina. Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web, 2005.

What is Web Spamming?

Example

Query: Kaiser Pharmacy (at the time paper was written 2005)

Result: had "*techdictionary.com*" as a 3rd hit

Why is Web Spamming Bad?

Search Engines suffer because

- It damages the **search engine's reputation**
- there is a the **cost** involved to crawling, indexing, storing the spam pages.

Users suffer because

- precision in the query results is lowered

How is it done?

Web Spamming Techniques involve

- **Boosting Techniques**
  - **Content** spam (e.g, Term repetition)
  - **Link** spam (e.g Link farming)
- **Hiding Techniques**

  For example, Content hiding, Cloaking

What do the Search Engines Ultimately want ?

- Want to calculate and return the exact page ranks based on relevance and importance

- Want to avoid the spam pages altogether before they use resources that might be used in storing or processing those web pages

The Agenda

- Focus of the First Paper
- Motivation
- **The Objective**
- DATASET
- Link Based Features
- Content Based Features
- Using the Content & Link Based Features
- Using the Web Topology
- Conclusion

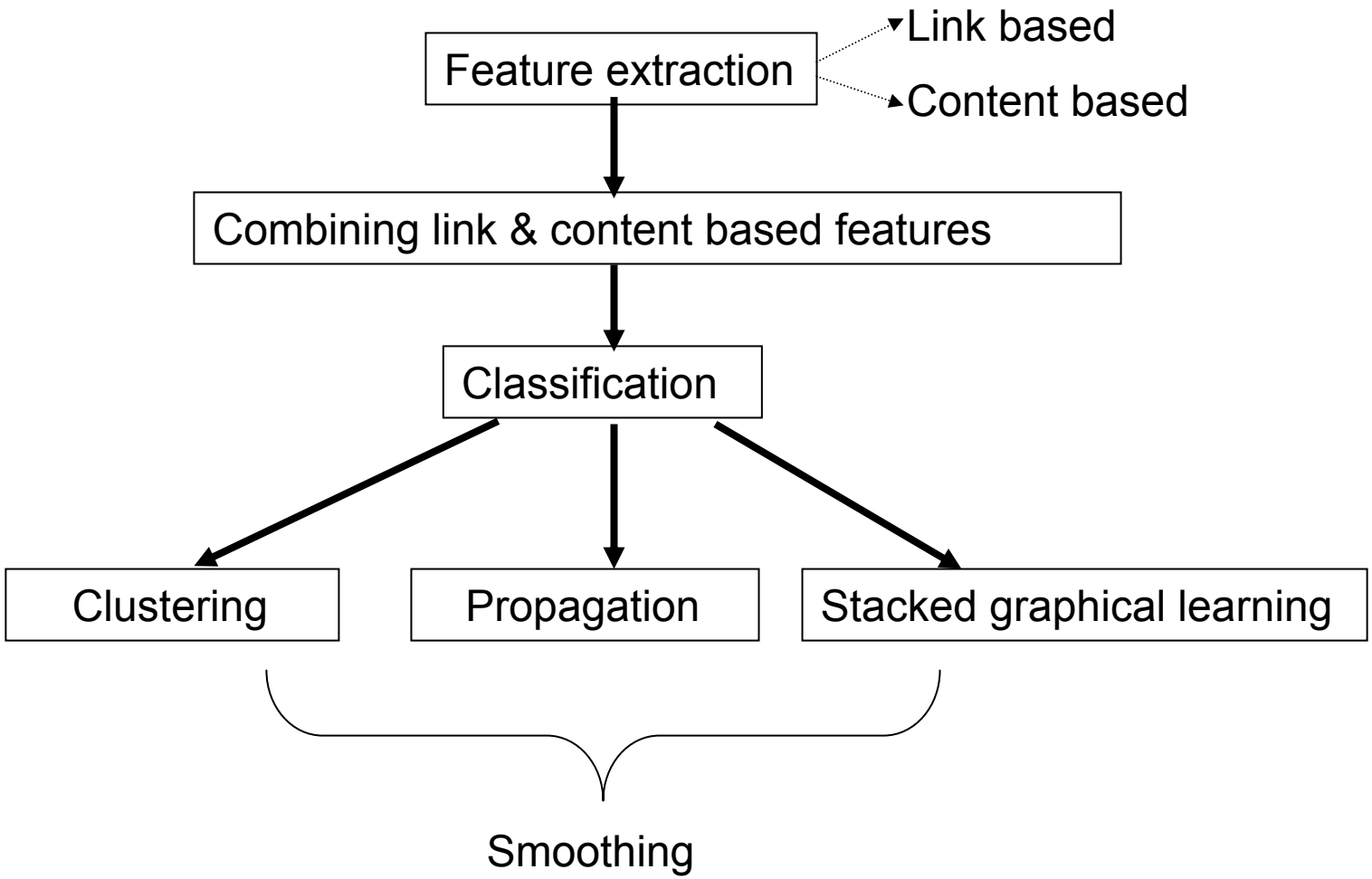
The OBJECTIVE

A Web Spam Detection System that is most accurate and reliable

The paper proposed a **Web Spam Detection System** that

1. uses the topology of the web graph by exploiting **dependencies among the web pages**
2. that **combines both the link and content based features**

The Flow


Feature extraction
Link based
Content based
Combining link & content based features
Classification
Clustering
Propagation
Stacked graphical learning
Smoothing

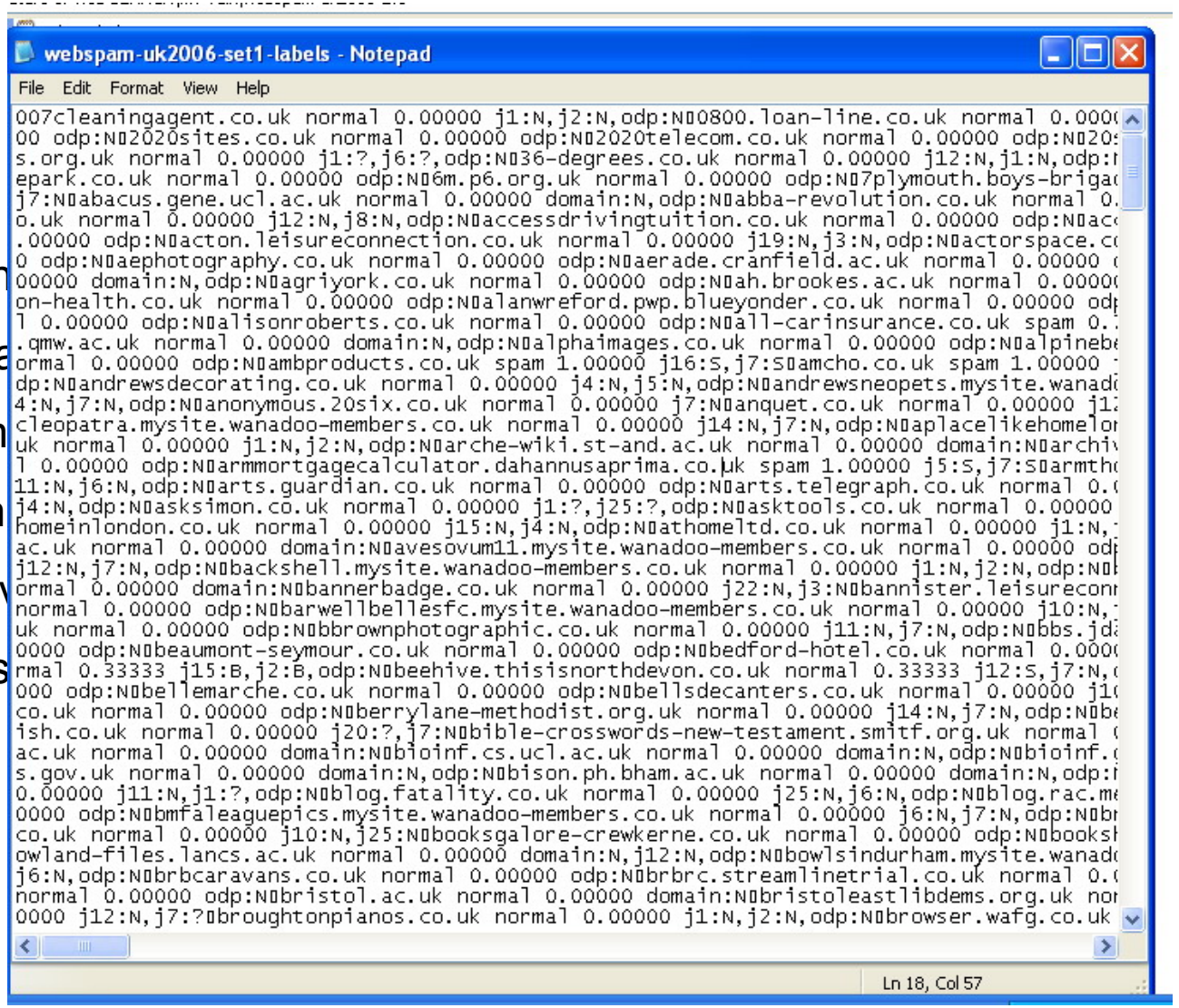
The Agenda

- Focus of the First Paper
- Motivation
- The Objective
- **DATASET**
- Link Based Features
- Content Based Features
- Using the Content & Link Based Features
- Using the Web Topology
- Conclusion

Data Set

Data Set

- Collected in
- Publicly ava
- Pages from
- 77.9 million
- A group of
- 6,552 hosts



[15] http://www.yr-bcn.es/webspam/datasets/

Distribution of host labels, as judged by human volunteers

Table 1: Distribution of host labels, as judged by human volunteers.

Label	Frequency	Percentage
Normal	4,046	61.75%
Spam	1,447	22.08%
Borderline	709	10.82%
Could not be classified	350	5.34%

Evaluation

- True positive rate, or Recall R
- False positive rate
- **F-measure**

The Agenda

- Focus of the First Paper
- Motivation
- The Objective
- DATASET
- **Link Based Features**
- Content Based Features
- Using the Content & Link Based Features
- Using the Web Topology
- Conclusion

Link Based Features

- Degree related measures
- PageRank
- TrustRank
- Truncated PageRank
- Estimation of d supporters

140 features per host

❖ Page Rank

Uses link structure to determine importance or popularity of a web page

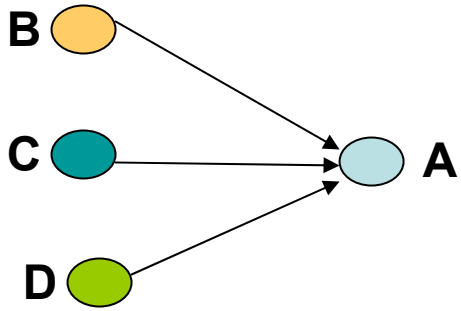
Intuition:

- A web page is important if several other important pages point to it
- PageRank of a page influences and is being influenced by importance of the other pages

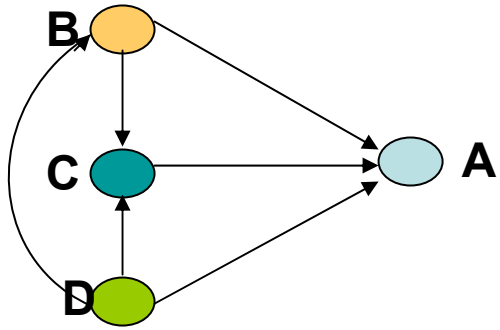
❖ Page Rank

Calculation: Initial scores already defined for all the pages

$$PR(A) = PR(B) + PR(C) + PR(D).$$



$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}.$$



[6] http://en.wikipedia.org/wiki/PageRank

❖ Page Rank

Page Rank with **Random Surfer Model**

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

d is damping factor,

$M(p_i)$ is the set of pages that link to *pi*,

$L(p_j)$ is the number of outbound links on page *pj*,

N is the total number of pages.

Link Based Features

❖ Trust Rank

A page having high pageRank is more likely to be spam if it had no relationship with a set of trusted pages

How it works?

- ❖ Determine the seed set **S** (using high PageRank or high out-degree)
- ❖ Determine "good" or "bad" nodes of the set **S** (using oracle)

$$O(p) = \begin{cases} 0 & \text{if } p \text{ is bad,} \\ 1 & \text{if } p \text{ is good.} \end{cases}$$

- ❖ E.g, S ={ ● , ● , ● }

where, good = ●

bad = ●

❖ Trust Rank

❖Calculate & propagate the trust from good pages (adjusting trust attenuation)

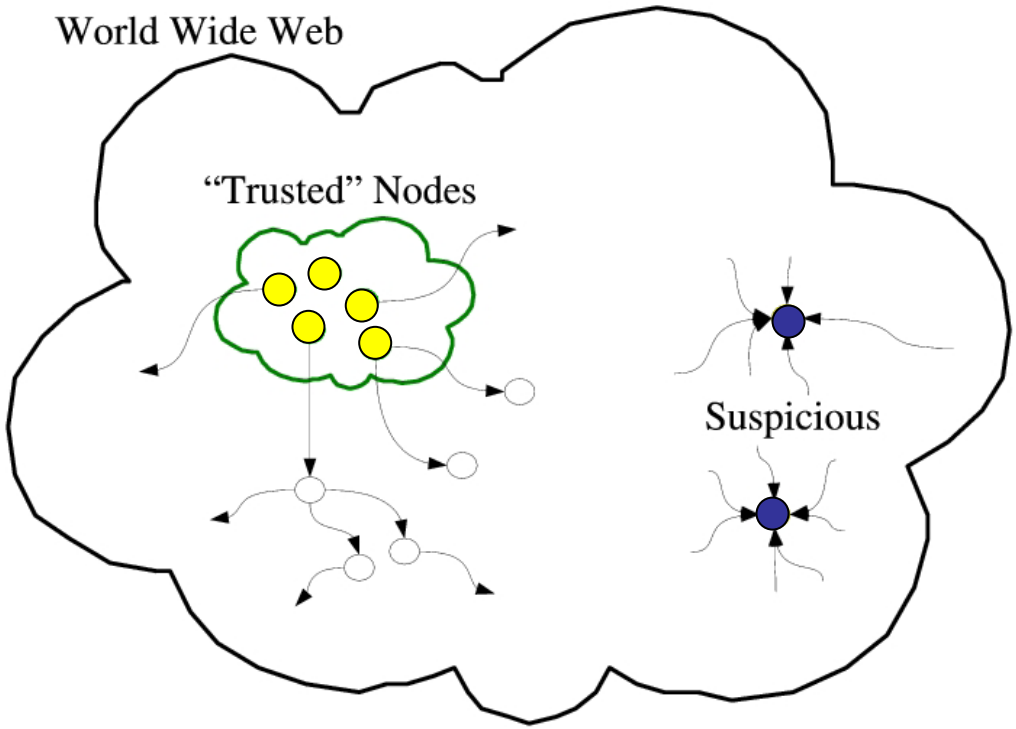


Image [4]: http://www.tejedoresdelweb.com/slides/2007_ojobuscador_madrid_spam.pdf

Histogram of ratio b/w TrustRank & PageRank

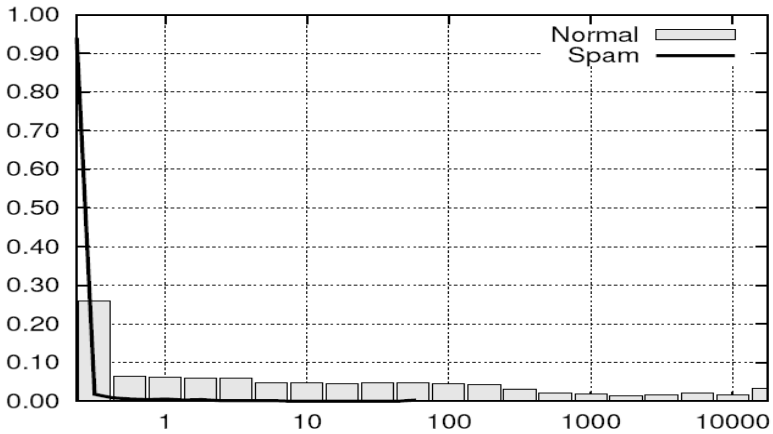
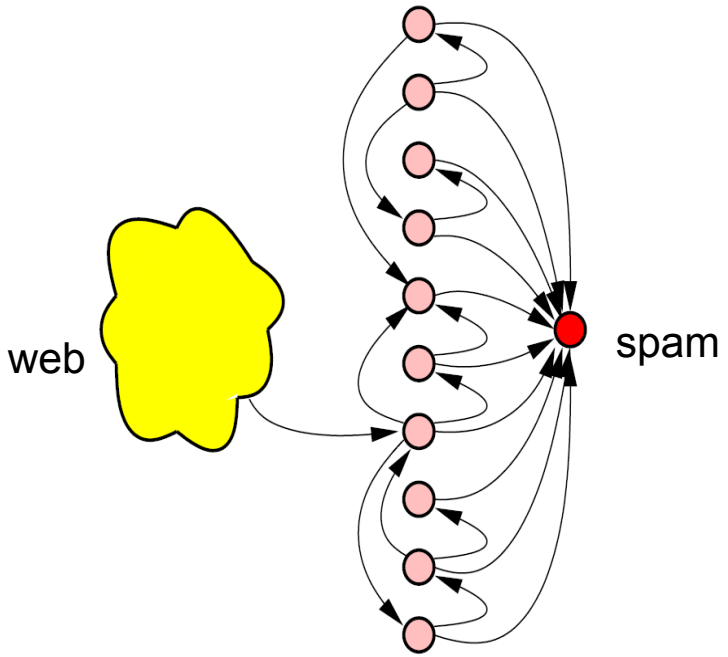
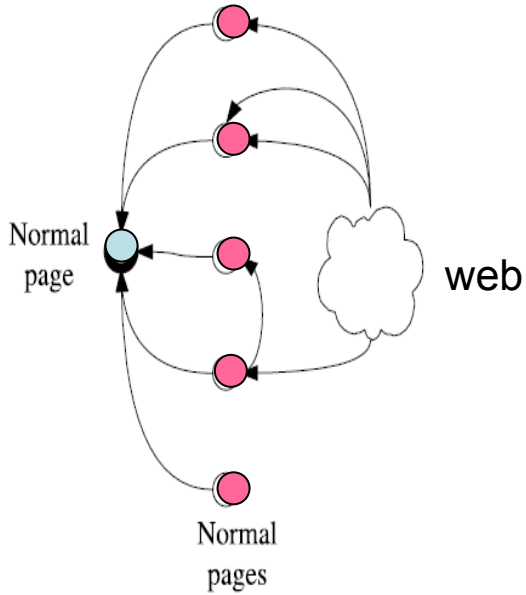


Figure 2: Histogram of the ratio between TrustRank and PageRank in the home pages.

❖ Truncated Page Rank

➢ A variant of PageRank, diminishes the influence of a page to the PageRank score of its close neighbors





❖ Estimation of d-supporters

- x is d-supporter of node y if shortest path from x to y has length ‘d’
- $N_d(x)$ be the set of d-supporters of page x
- For each page x, cardinality of $N_d(x)$ is an increasing fuction with respect to d.

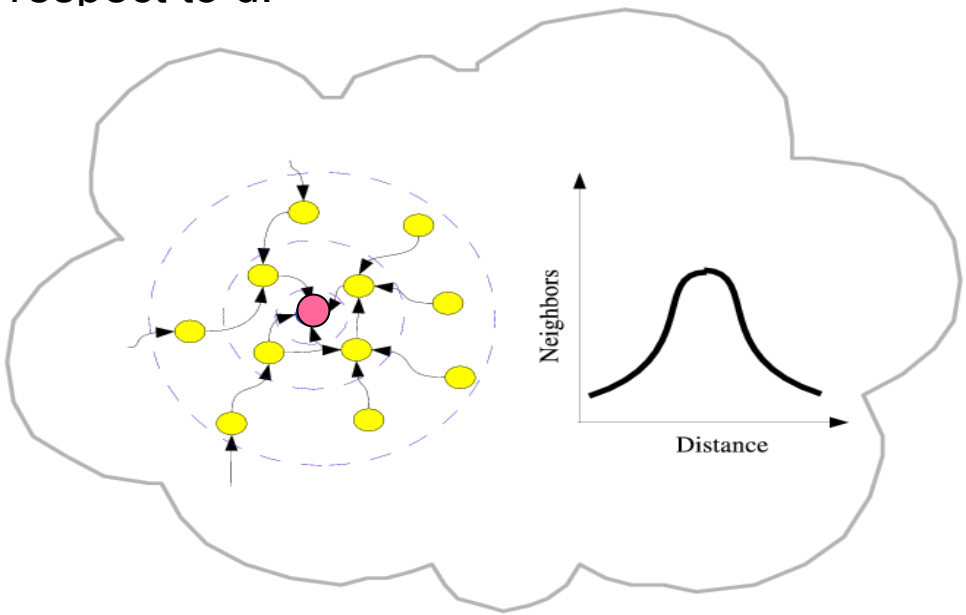


Image [4]

Bottleneck Number

The bottleneck measure for page x, defined as

$$b_d(x) = \min_{j \leq d}\{|N_j(x)|/|N_{j-1}(x)|\}$$

- indicates the minimum rate of growth of neighbors of x up to a certain distance
- spam pages have smaller bottle neck numbers than non-spam pages

Bottleneck: Non-Spam & Spam

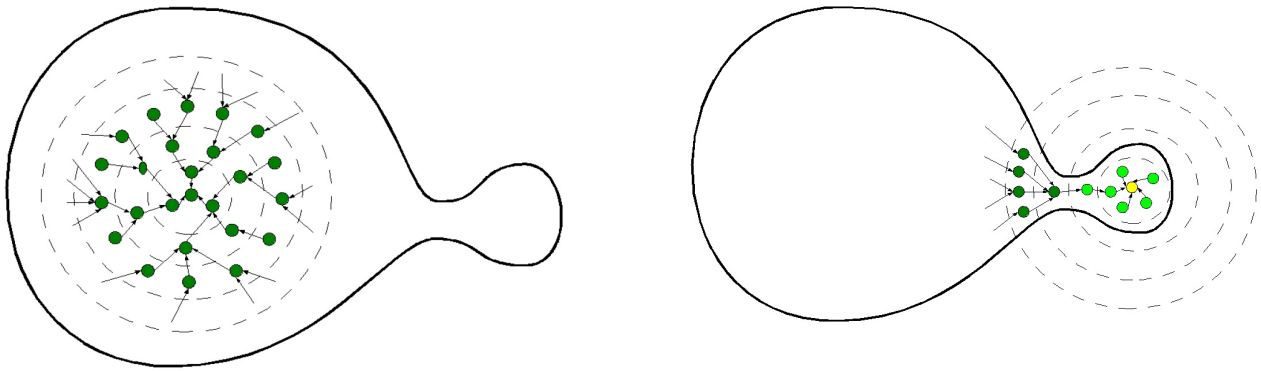


Image [4]

Histogram of $b_4(x)$ of Spam and Non-spam Pages

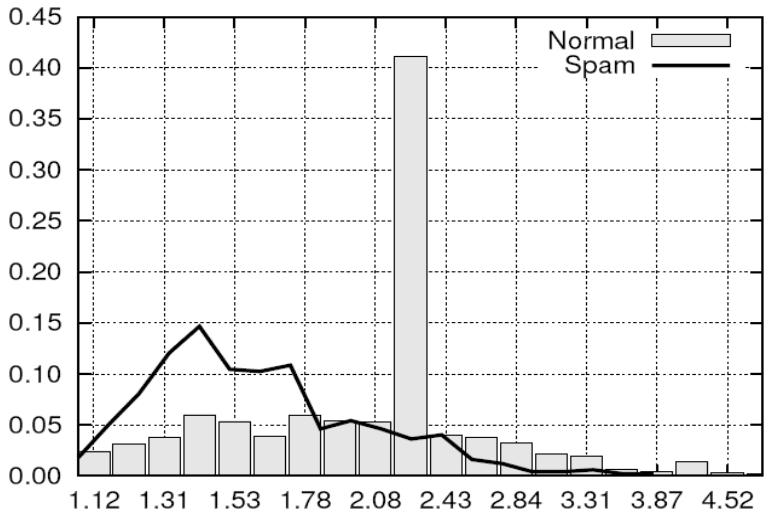


Figure 3: Histogram of the minimum ratio change of the # of neighbors from distance i to distance $i-1$

The Agenda

- Focus of the First Paper
- Motivation
- The Objective
- DATASET
- Link Based Features
- **Content Based Features**
- Using the Content & Link Based Features
- Using the Web Topology
- Conclusion

Content Based Features

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

96 features per host

Content Based Features

- **Average word length**

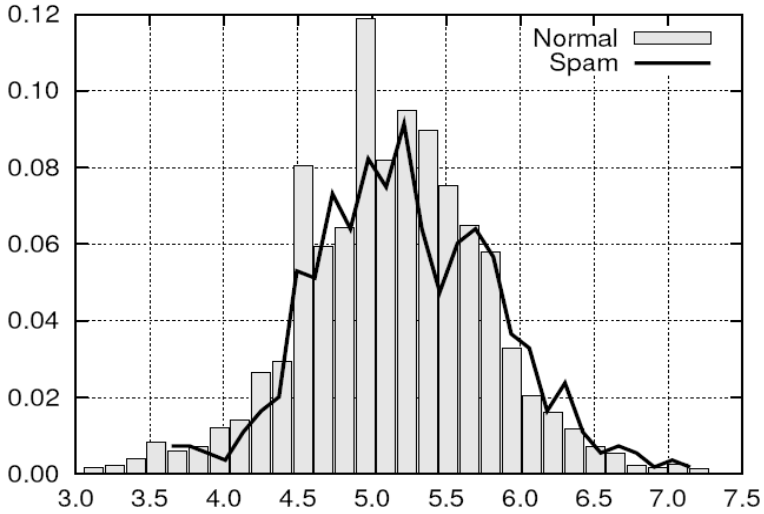


Figure 4: Histogram of the average word length in non-spam vs. spam pages.

Content Based Features

❖ Compression Rate

$$\text{Compression rate} = \frac{\text{size of compressed text (visible text)}}{\text{size of uncompressed text}}$$

❖ Precision and Recall

F= Frequent terms in the collection

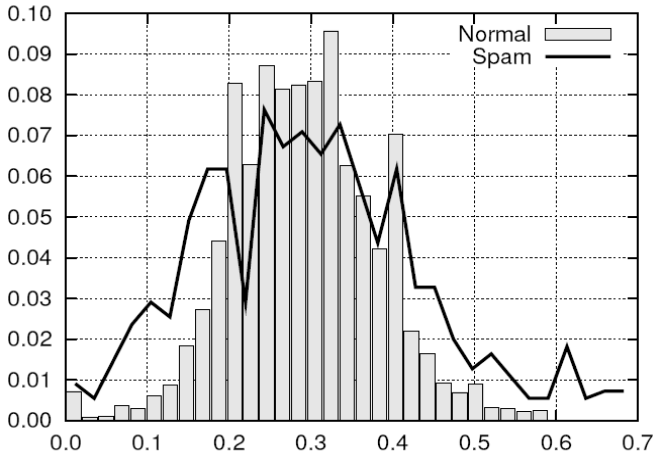
T= Terms in the page

Q= Frequent terms in the query log

Corpus Recall = $|F \cap T| \; / \; |F|$

Query Recall = $|Q \cap T| \; / \; |Q|$

Content Based Features



Corpus Precision = | F ∩ T | / | T |

Figure 5: Histogram of the corpus precision in non-spam vs. spam pages.

Query Precision = | Q ∩ T | / | T |

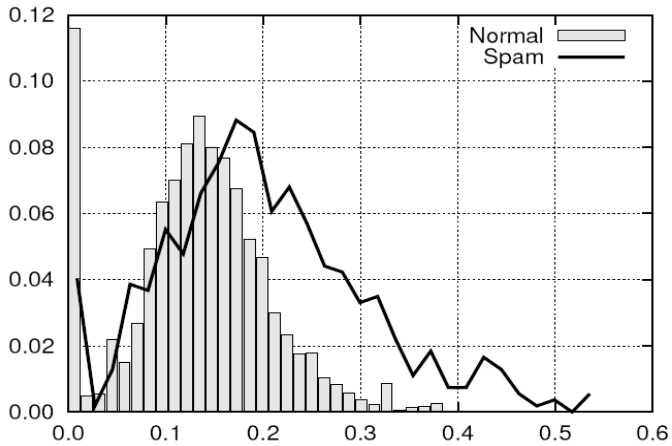


Figure 6: Histogram of the query precision in non-spam vs. spam pages for $k = 500$.

Content Based Features

- **Entropy of trigrams (Compression)**
  - Calculated on distribution of trigrams
  - Let

    $\{p_w\}$ be probability distribution on trigrams of a page

    $T = \{w\}$ be the set of all trigrams in a page

Entropy of trigrams= $H = -\sum_{w \in T} p_w \log p_w$

The Agenda

- Focus of the First Paper
- Motivation
- The Objective
- DATASET
- Link Based Features
- Content Based Features
- **Using the Link and Content Features**
- Using the Web Topology
- Conclusion

Cost Sensitive Decision Tree

Table 2: Cost-sensitive decision tree

Cost ratio (R)	1	10	20	30	50
True positive rate	64.0%	68.0%	75.6%	80.1%	87.0%
False positive rate	5.6%	6.8%	8.5%	10.7%	15.4%
F-Measure	0.632	0.633	**0.646**	0.642	0.594

Bagging

Table 3: Bagging with a cost-sensitive decision tree

Cost ratio (R)	1	10	20	30	50
True positive rate	65.8%	66.7%	71.1%	78.7%	84.1%
False positive rate	2.8%	3.4%	4.5%	5.7%	8.6%
F-Measure	0.712	0.703	0.704	**0.723**	0.692

Using Content & Link Based Features

Comparing Link and Content based features

Table 4: Comparing link and content features

	Both	Link-only	Content-only
True positive rate	78.7%	79.4%	64.9%
False positive rate	5.7%	9.0%	3.7%
F-Measure	**0.723**	0.659	0.683

The Agenda

- Focus of the First Paper
- Motivation
- The Objective
- DATASET
- Link Based Features
- Content Based Features
- Using the Content & Link Based Features
- **Using the Web Topology**
- Conclusion

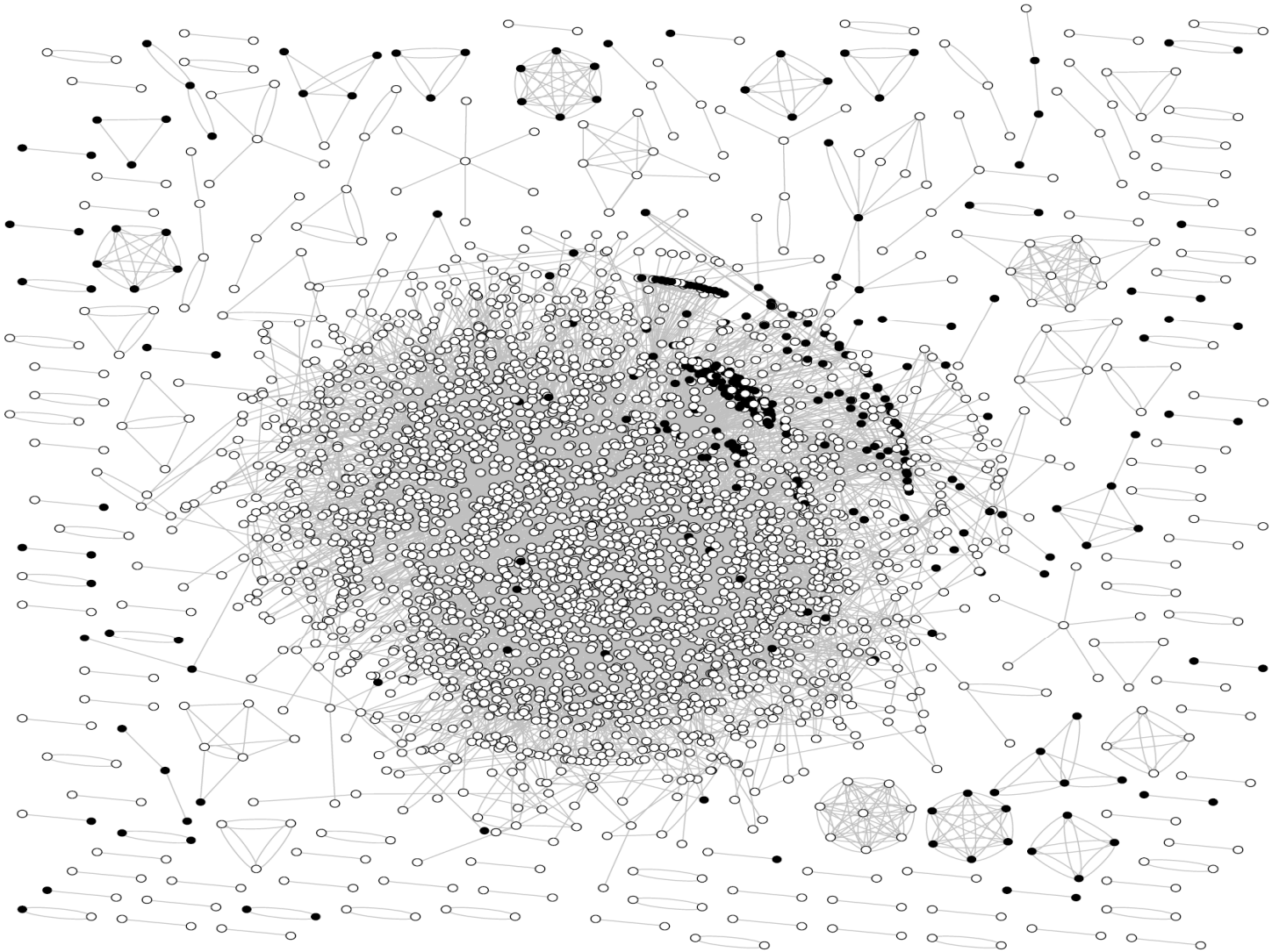
Using the Web Topology

Observation :

Similar pages tend to be linked to be linked together more frequently than dissimilar ones

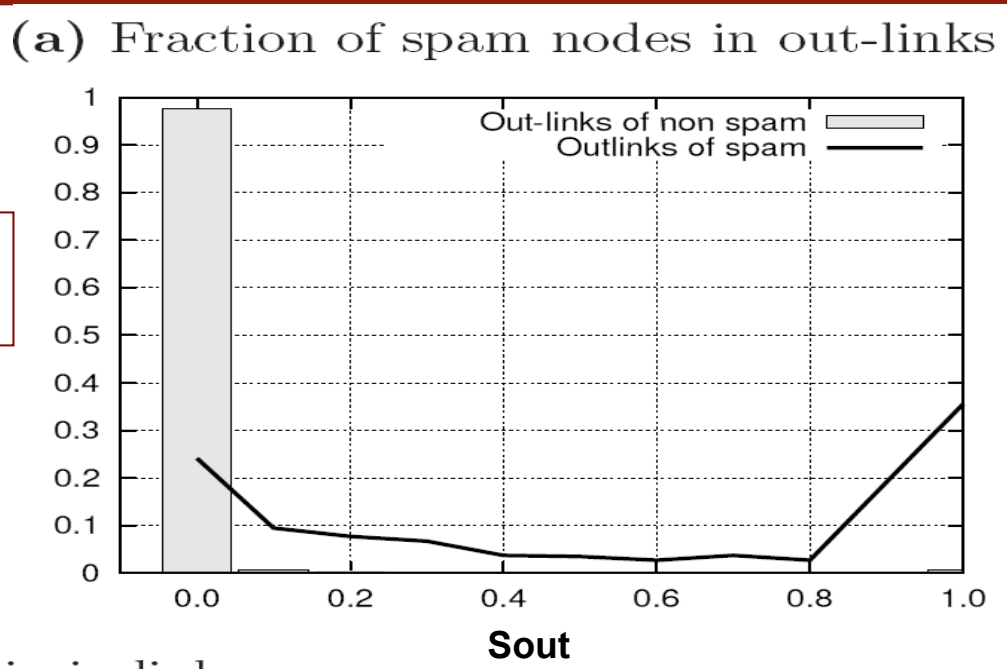
Similar pages tend to be linked to be linked together more frequently than dissimilar ones

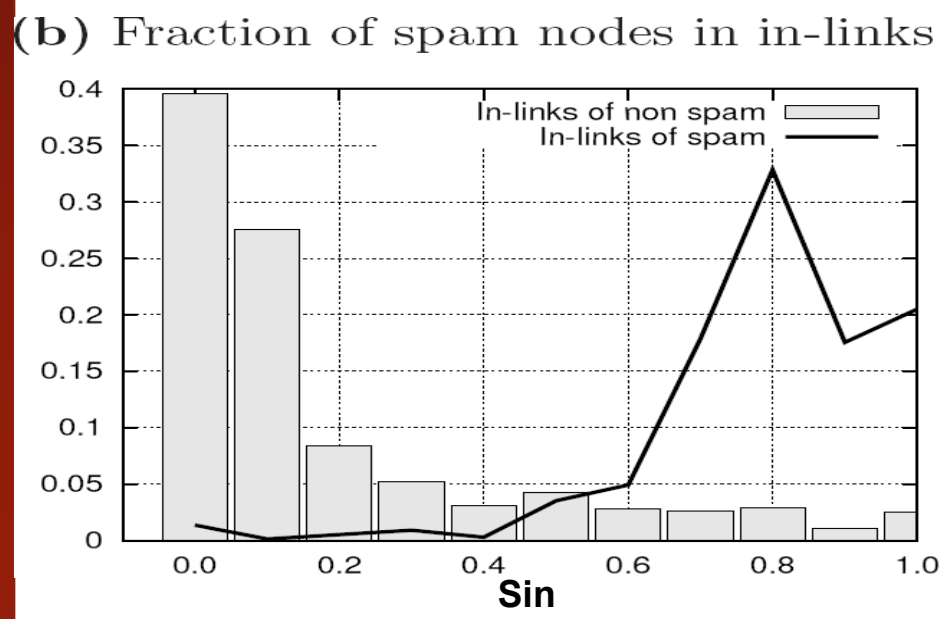
Using the Web Topology



Using the Web Topology -Topological Dependencies of Spam Nodes

Sout= No. of spam hosts linked by x / All hosts linked by x

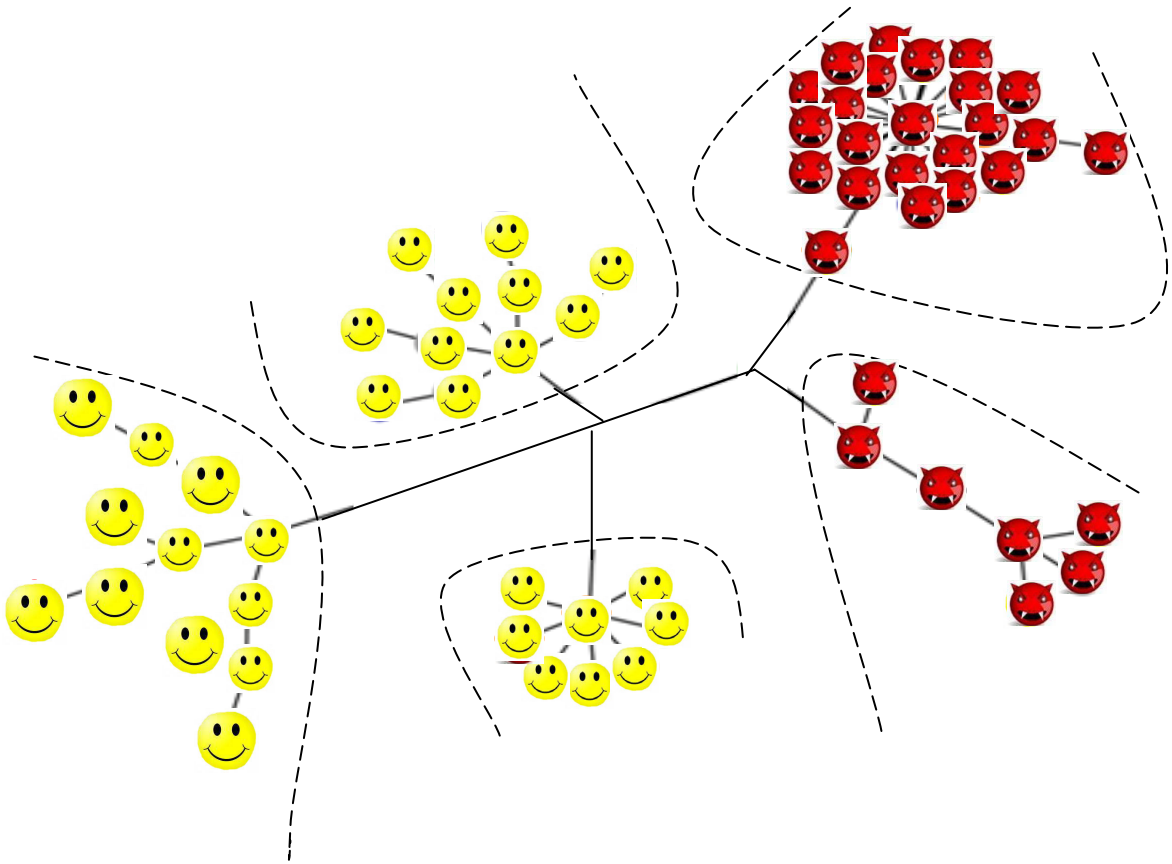




Sin= No. of spam hosts linked to x / All hosts linked to x

Clustering

Using METIS graph clustering algorithm



Clustering

Table 5: Results of the clustering algorithm

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

Propagation

Use the graph topology to smooth predictions by propagating them as *random walks*

Main Idea

Use the *predicted spamicity* of a particular classification method and start a random walk with the restart probabilty 1-α

$$\mathbf{v}_h^{(t+1)} = (1-\alpha)\mathbf{v}^{(0)} + \alpha \sum_{g:g\to h} \frac{\mathbf{v}_g^{(t)}}{\text{outdeg}(g)}$$

Where

h: host

p(h) : [0..1] (p(h)=0: *non-spam* , p(h)=1: *spam*, for each host h)

v(0) : vector such that $v^{(0)}_h = p(h)/\sum_{h\in H} p(h)$

outdeg(g): the out-degree of g

Propagation

- Applied with three forms of random walk
  - on the host graph (forward)
  - on the transposed host graph (backward)
  - on the undirected host graph (both)
- Observed improvements on $\alpha \in [0.1, 0.3]$ out of other tried values

Propagation

Results with $\alpha = 0.3$ in the table

Table 6: Result of applying propagation

	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	**0.676**
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

Stacked Graph Learning

- Uses initial predictions $p(h) \in [0..1]$ by the classification scheme C for all the objects in the data set
- Generates a set of extra features for each object based on relevance (w.r.t in-links, outlinks or both)
  - Let r(h) be the set if pages related to host h
  - Then, $f(h) = \frac{\sum_{g \in r(h)} p(g)}{|r(h)|}$
- Adds these extra features f(h) to the input of C and run the algorithm again

Stacked Graph Learning

Improvement is seen in the first Iteration

Table 7: Results with stacked graphical learning

	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	**0.750**

Stacked Graph Learning

Second Iteration showed even better results

Table 8: Second pass of stacked graphical learning

	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	**0.763**

The Agenda

- Focus of the First Paper
- Motivation
- The Objective
- DATASET
- Link Based Features
- Content Based Features
- Using the Web Topology
- **Conclusion**

Conclusion

Experiments have clearly shown that

- Combining both the link and content based features brings significant improvement to spam detection techniques
- Using web graph dependencies , we can detect the web spam by turning spammers' ingenuity against themselves

Thank You!

Q&A and Discussion

References-1

[1] www.milliondollarhomepage.com

[2] http://www.dcc.uchile.cl/~ccastill/papers/cdgms_2006_know_your_neighbors.pdf

[3] http://images.google.com/imgres?

[4] http://www.tejedoresdelweb.com/slides/2007_ojobuscador_madrid_spam.pdf

[5] Z. Gyongyi and H. Garcia-Molina. Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web, 2005.

[6] http://en.wikipedia.org/wiki/PageRank

[7] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Link-based characterization and detection of Web Spam. In AIRWeb, 2006.

[8] http://infolab.stanford.edu/~zoltan/publications/gyongyi2006link.pdf

References-2

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

[10] http://en.wikipedia.org/wiki/Cross-validation

[11] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In Proceedings of the World Wide Web conference, pages 83–92, Edinburgh, Scotland, May 2006.

[12] http://www.salford-systems.com/doc/BAGGING_PREDICTORS.pdf

[13] Z. Gy¨ongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In Proc. of the 30th VLDB Conf., 2004.

[14] M. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. ACM SIGIR Forum, 36(2), 2002.

[15] http://www.yr-bcn.es/webspam/datasets/