



HY463 - Συστήματα Ανάκτησης Πληροφοριών  
Information Retrieval (IR) Systems

## Web Searching III

### Web Spam Page Identification Εντοπισμός «Παραπλανητικών» Ιστοσελίδων

Γιάννης Τζιζίκας

Διάλεξη : 10

Ημερομηνία :

Based on

Z. Gyongyi, H. Garcia-Molina, J. Pedersen,  
Computing Web Spam with Trust Rank, SIGMOD'04



## Κίνητρο

- Υπάρχουν πολλές ιστοσελίδες (και καθημερινά δημιουργούνται νέες) οι οποίες χρησιμοποιούν διάφορα τεχνάσματα ώστε να εξαπατήσουν τις μηχανές αναζήτησης και να λάβουν βαθμό (π.χ. PageRank) υψηλότερο από αυτό που «αξίζουν», και άρα να εμφανίζονται σε υψηλές θέσεις στα αποτελέσματα επερωτήσεων.
- Ο εντοπισμός των σπαμ σελίδων από ανθρώπους είναι εφικτός αλλά αυτό θα ήταν εξαιρετικά χρονοβόρο και ακριβό για μεγάλο πλήθος σελίδων
- Ανάγκη για αυτόματες ή ημιαυτόματες τεχνικές διαχωρισμού των «καλών» σελίδων από τις «κακές»

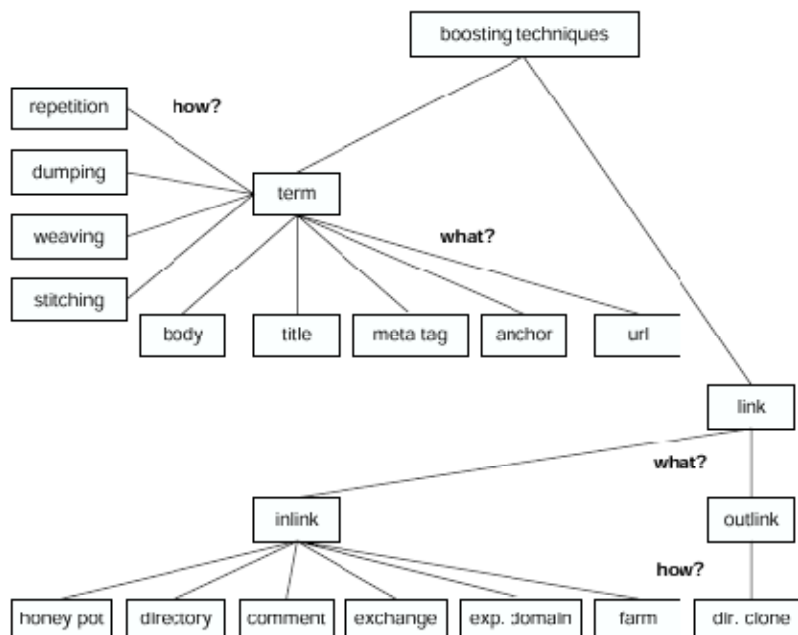


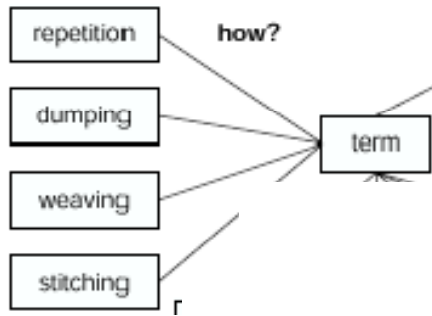
## Ορισμός: διασυνδεδεμένες σελίδες δημιουργημένες για παραπλάνηση των μηχανών αναζήτησης

- Παραδείγματα
- a pornography site page that contains thousands of keywords which are made invisible (to humans) by adjusting accordingly the color scheme
  - a search engine will include this page in the results of a query that contains some of these keywords
- creation of a large number of bogus web pages, all pointing to a single target page (that page will have high in-degree)
  - a search engine will rank high this page

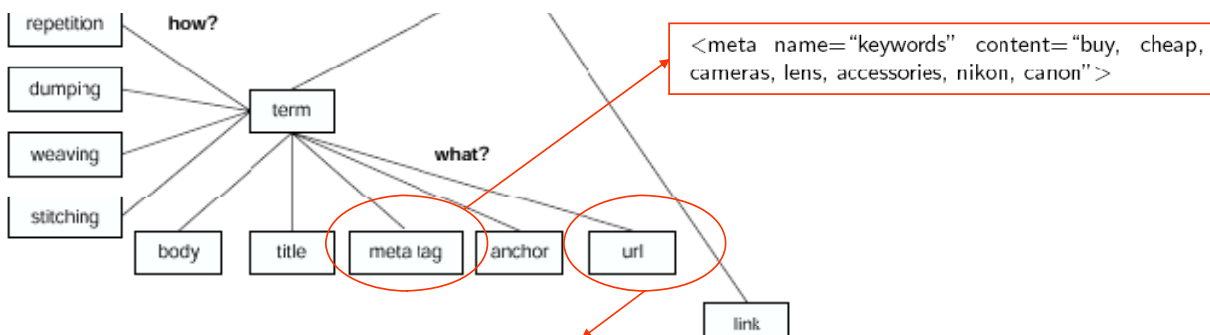


## Μια Ταξινόμηση των Τρόπων Παραπλάνησης Μηχανών Αναζήτησης Πηγή "Web Spam Taxonomy" (<http://airweb.cse.lehigh.edu/2005/gyongyi.pdf>)





- **Repetition** of one or a few specific terms.
  - They achieve increased relevance for a document with respect to a small number of query terms.
- **Dumping** of a large number of unrelated terms, often even entire dictionaries.
  - They achieve making a certain page relevant to many different queries. Effective against queries that include relatively rare, obscure terms: for such queries, it is probable that only a couple of pages are relevant, so even a spam page with a low relevance/importance would appear among the top results.
- **Weaving** of spam terms into copied contents.
  - spammers duplicate text corpora (e.g., news articles) available on the Web and insert spam terms into them at random positions. Effective if the topic of the original real text was so rare that only a small number of relevant pages exist.
- **Phrase stitching**: they glue together sentences or phrases, possibly from different sources;
  - the spam page might then show up for queries on any of the topics of the original sentences.



### URL spam.

Some search engines also break down the URL of a page into a set of terms that are used to determine the relevance of the page. To exploit this, spammers sometimes create long URLs that include sequences of spam terms. For instance, one could encounter spam URLs like:

- buy-canon-rebel-20d-lens-case.camerasx.com,
- buy-nikon-d100-d70-lens-case.camerasx.com,
- Some spammers even go to the extent of setting up a DNS server that resolves any host name within a domain.



## Spam Hiding Techniques

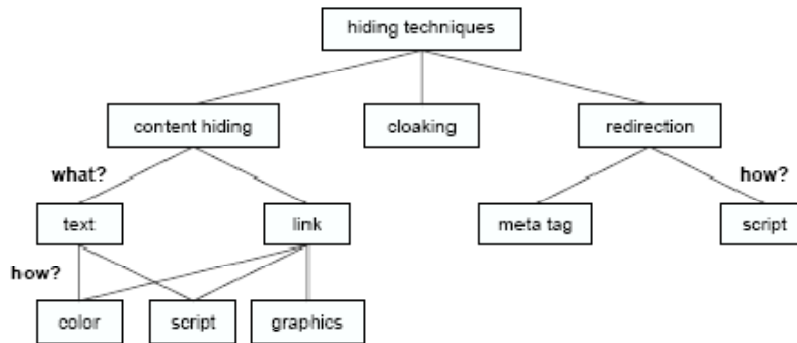


Figure 3: Spam hiding techniques.

```

<body background="white">
  <font color="white">hidden text</font>
  ...
</body>

```

Hidden text

```

<a href="target.html"></a>

```

Hidden link



## Spam Hiding Techniques > Cloaking

- Given a URL, spam web servers return one specific HTML document to a regular web browser, while they return a different document to a web crawler.
- How?
  - (a) Spammers can maintain a list of IP addresses used by search engines, and identify web crawlers based on their matching IPs.
  - (b) A web server can identify the application requesting a document based on the user-agent field in the HTTP request message, e.g.

GET /db pages/members.html HTTP/1.0

Host: www-db.stanford.edu

User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)



### Κλασσικός τρόπος αντιμετώπισης

- Οι εταιρίες που διατηρούν μηχανές αναζήτησης προσλαμβάνουν προσωπικό «ειδικευμένο» στον εντοπισμό σπαμ. Το προσωπικό αυτό συνεχώς σαρώνει τον ιστό για τον εντοπισμό «κακόβουλων» σελίδων.
- Αν μια παραπλανητική σελίδα (ή site) εντοπιστεί, τότε η μηχανή παύει να την ευρετηριάζει (μπαίνει στη μαύρη λίστα του ερπυστή και διαγράφεται από το ευρετήριο της μηχανής).

### Πολύ δαπανηρή και αργή διαδικασία



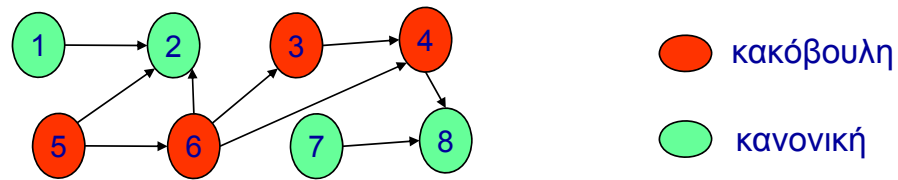
- 1) Επιλογή ενός μικρού συνόλου από σελίδες σπόρους (seed pages) για αξιολόγηση από ειδικούς (experts)
- 2) Μετά τη χειρονακτική επιλογή των αξιόπιστων σελίδων σπόρων, η δομή του γράφου του Ιστού μπορεί να αξιοποιηθεί προκειμένου να ανακαλύψουμε άλλες σελίδες που *πιθανώς είναι επίσης καλές*.

### Ζητήματα:

- Πως να επιλέξουμε το σύνολο των σπόρων (seed selection);
- Πως να ανακαλύψουμε τις καλές σελίδες;



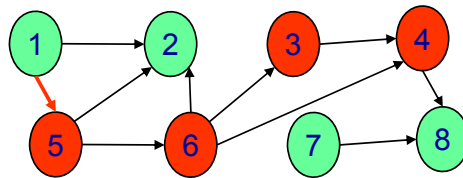
## Εμπειρική Παρατήρηση: απομόνωση των καλών σελίδων (Approximate isolation of the good set)



- **Εμπειρική παρατήρηση:**
  - Οι καλές σελίδες σπάνια δείχνουν σε κακές



## Εμπειρική Παρατήρηση: απομόνωση των καλών σελίδων Εξαιρέσεις



the creators of good pages can sometimes be «tricked» and add links to bad pages.

Examples:

- Unmoderated message boards where spammers post messages that include links to their spam pages
- **Honey pots**
  - pages that contain **some useful resource** but have hidden links to their spam pages (the honey pot attracts people to point to it)



- We can formalize the notion of a human checking a page by a binary «oracle» function  $O$ , over all pages  $p$  in  $W$ .

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

- Oracle invocations are expensive and time consuming. We do not want to call the oracle function for all pages. Our objective is to be selective, i.e. to ask a human expert to evaluate only some of the pages



## Συνάρτηση Εμπιστοσύνης (Trust Function)

- To evaluate pages without relying on  $O$ , we will estimate the likelihood that a given page  $p$  is good. To this end we will introduce *trust functions*.
- Trust function yields values between 0 (**bad**) and 1 (**good**)

Στην ουσία θέλουμε από την συνάρτηση  $O$  που έχει πεδίο ορισμού τις σελίδες σπόρους, να ορίσουμε μια συνάρτηση εμπιστοσύνης  $T$  η οποία να έχει πεδίο ορισμού το σύνολο όλων των σελίδων.

Άρα από την

$$O: S \rightarrow [0,1]$$

να ορίσουμε μια

$$T: W \rightarrow [0,1]$$

όπου

- $W$ : the set of all pages
- $S$ : the set of seed pages
  - $S \subset W$ , και  $|S| \ll |W|$  (δηλαδή το  $S$  είναι πολύ μικρότερο του  $W$ )



## Συνάρτηση Εμπιστοσύνης (Trust Function)

- Ideally, for any  $p$ ,  $T(p)$  should give us the probability that  $p$  is good
- Ideal Trust Property (ITP)
  - $T(p) = \Pr[O(p)=1]$ 
    - // δηλαδή  $T(p)$  = πιθανότητα η μαντική συνάρτηση να μας δώσει 1
  - difficult to achieve
  - even if  $T$  is not very accurate we could exploit it to order pages by their likelihood of being good
- Desired Trust Property (relaxation of ITP)
  - $T(p) < T(q) \Leftrightarrow \Pr[O(p)=1] < \Pr[O(q)=1]$
  - $T(p) = T(q) \Leftrightarrow \Pr[O(p)=1] = \Pr[O(q)=1]$
- Threshold Trust Property (another relaxation of ITP)
  - $T(p) > \delta \Leftrightarrow O(p)=1$



## Υπολογισμός Εμπιστοσύνης: The ignorant trust function

### The ignorant trust function $T_0$

- We can select at random a seed set  $S$  of  $L$  pages and call the oracle on its elements.
- Let  $S^+$  be the good pages and  $S^-$  the bad ones. Since the remaining pages are not checked we can mark them with  $1/2$ .
- We can call this ignorant trust function  $T_0$

$$T_0(p) = \begin{cases} O(p) & \text{if } p \in S \\ 1/2 & \text{otherwise} \end{cases}$$





## Διάδοση Εμπιστοσύνης (Trust Propagation)

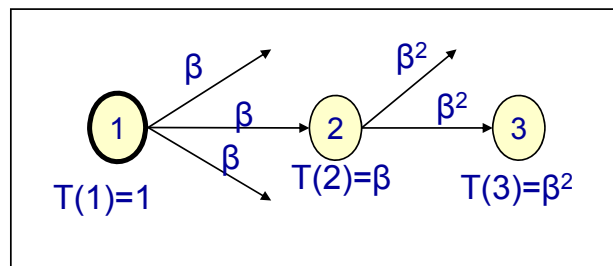
- We can exploit the empirical observation «**Good pages seldom point to bad ones**», and assign score 1 to all pages that are reachable from a page in  $S^+$  in  $M$  or fewer steps.
- Trust Function  $T_M$ :

$$T_M(p) = \begin{cases} O(p) & \text{if } p \in S \\ 1 & \text{if } p \notin S \text{ and } \exists q \in S^+ : q \xrightarrow{M} p \\ 1/2 & \text{otherwise} \end{cases}$$

- The notation  $q \xrightarrow{M} p$  means : there is a path of maximum length  $M$  from  $q$  to  $p$
- The bigger  $M$  the further we are from good pages, the less certain we are that a page is good



## Εξασθένηση Εμπιστοσύνης (Trust Attenuation) Trust dampening

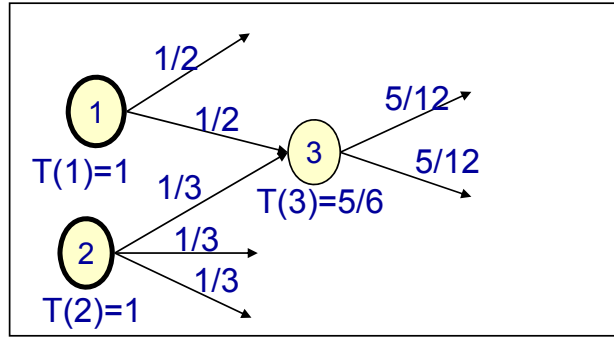


### Trust dampening

- assign a score  $\beta$  ( $<1$ ) to pages reachable at 1 step
- assign the score  $\beta^2$  to pages reachable at 2 step, and so on
- pages with multiple inlinks: maximum score or average score



## Εξασθένηση Εμπιστοσύνης (Trust Attenuation) Trust splitting



### Trust splitting

- motivation: the care with which people add links to their pages is often inversely proportional to the number of links on the page
  - if page  $p$  has a trust score  $T(p)$  and it points to  $|\text{out}(p)|$  pages, each of them will receive a score fraction  $T(p)/|\text{out}(p)|$  from  $p$
  - the actual score of a page will be the sum of the score fractions received through its inlinks
- We could combine trust *dampening* and *splitting*

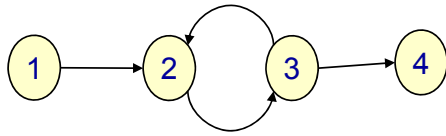


## Ο Αλγόριθμος TrustRank

- In TrustRank we will combine **trust dampening** and **splitting**:
  - in each iteration, the trust score of a node is split among its neighbors and dampened by a factor of  $a_b$
- We will compute TrustRank scores using a biased PageRank algorithm
  - the oracle-provided scores replace the uniform distribution
    - In PageRank we were using the uniform distribution to express the random jumps of the random surfer



## Επανάληψη: PageRank



Adjacency matrix  $M$

$$M = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Transition matrix  $T$

$$T(p, q) = \begin{cases} 0 & \text{if } (q, p) \notin M \\ 1/|\text{out}(q)| & \text{if } (q, p) \in M \end{cases}$$

$$T = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1/2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \end{pmatrix}$$

- The PageRank score  $R(p)$  of a page is defined as

$$R(p) = a \cdot \sum_{q \in \text{in}(p)} \frac{R(q)}{|\text{out}(q)|} + (1-a) \frac{1}{N}$$

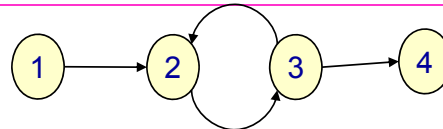
- The equivalent matrix equation:

$$R = a \cdot T \cdot R + (1-a) \frac{1}{N} \mathbf{1}_N$$



## Επανάληψη: PageRank

$$R = a \cdot T \cdot R + (1-a) \frac{1}{N} \mathbf{1}_N$$



$$\begin{bmatrix} r1 \\ r2 \\ r3 \\ r4 \end{bmatrix} = a \cdot \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1/2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \end{bmatrix} \cdot \begin{bmatrix} r1 \\ r2 \\ r3 \\ r4 \end{bmatrix} + (1-a) \frac{1}{4} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} r1 \\ r2 \\ r3 \\ r4 \end{bmatrix} = a \cdot \begin{bmatrix} 0 \\ r1+r3/2 \\ r2 \\ r3/2 \end{bmatrix} + (1-a) \frac{1}{4} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} r1 \\ r2 \\ r3 \\ r4 \end{bmatrix} = \begin{bmatrix} (1-a)/4 \\ a(r1+r3/2)+(1-a)/4 \\ ar2+(1-a)/4 \\ ar3/2+(1-a)/4 \end{bmatrix}$$



## Επανάληψη: Ο Αλγόριθμος PageRank

function **PageRank**

Input T: transition matrix, N: number of pages,  
a<sub>b</sub>: decay factor for biased PageRank, M<sub>b</sub>: number of biased PageRank iterations  
output t\* : PageRank scores

(3)  $\mathbf{d} = 1/N * \mathbf{1}_N$  // initial score for all pages is 1/N

(5)  $\mathbf{t}^* = \mathbf{d}$   
for i=1 to M<sub>b</sub> do // evaluates PageRank scores  
     $\mathbf{t}^* = a_b T \mathbf{t}^* + (1 - a_b) \mathbf{d}$   
return  $\mathbf{t}^*$



## Ο Αλγόριθμος TrustRank

function **TrustRank**

Input T: transition matrix, N: number of pages, L: limit of oracle invocations,  
a<sub>b</sub>: decay factor for biased PageRank, M<sub>b</sub>: number of biased PageRank iterations  
output t\* : TrustRank scores

(1)  $\mathbf{s} = \text{SelectSeed}()$  // seed-desirability: returns a vector.

// E.g.  $\mathbf{s}(p)$  is the desirability for page p

(2)  $\sigma = \text{Rank}(\{1, \dots, N\}, \mathbf{s})$  // orders in decreasing order of s-value all pages

(3)  $\mathbf{d} = \mathbf{0}_N$  // initial score for all pages is 0

for i=1 to L do // invokes oracle function on the most desirable pages

    if  $O(\sigma(i)) = 1$  then  $\mathbf{d}(\sigma(i)) = 1$

(4)  $\mathbf{d} := \mathbf{d} / |\mathbf{d}|$  // normalize static distribution score (to sum up to 1)

(5)  $\mathbf{t}^* = \mathbf{d}$

for i=1 to M<sub>b</sub> do // evaluates TrustRank scores using a biased PageRank

$\mathbf{t}^* = a_b T \mathbf{t}^* + (1 - a_b) \mathbf{d}$  // note that  $\mathbf{d}$  replaces the uniform distribution

return  $\mathbf{t}^*$



## Ο Αλγόριθμος TrustRank

### Remarks:

- Step 5 implements a particular version of **trust dampening** and **splitting**: in each iteration, the trust score of a node is split among its neighbors and dampened by a factor of  $a_b$
- The good seed pages have no longer a score of 1, however they still have the highest scores

```

(4)  $\mathbf{d} := \mathbf{d} / |\mathbf{d}|$  // normalize static distribution score (to sum up to 1)
(5)  $\mathbf{t}^* = \mathbf{d}$ 
    for  $i=1$  to  $M_b$  do // evaluates TrustRank scores using a biased PageRank
         $\mathbf{t}^* = a_b \mathbf{T} \mathbf{t}^* + (1 - a_b) \mathbf{d}$  // note that  $\mathbf{d}$  replaces the uniform distribution
    return  $\mathbf{t}^*$ 

```



## Επιλέγοντας σπόρους Selecting Seeds

```

(1)  $\mathbf{s} = \text{SelectSeed}()$  // seed-desirability: returns a vector.
    // E.g.  $\mathbf{s}(p)$  is the desirability for page  $p$ 
(2)  $\sigma = \text{Rank}(\{1, \dots, N\}, \mathbf{s})$  // orders in decreasing order of s-value all pages

```

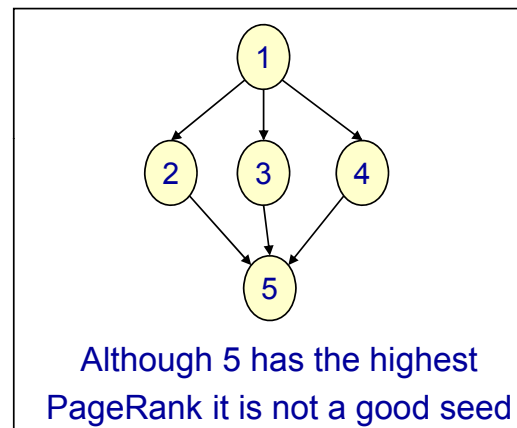
### Πιθανές Στρατηγικές

α) Random selection

β) **High PageRank**

Επιλέγουμε τις σελίδες με **υψηλό PageRank σκορ** διότι αυτές οι σελίδες συχνά εμφανίζονται στην κορυφή των απαντήσεων

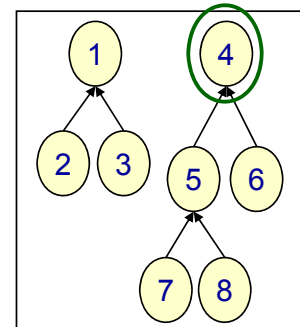
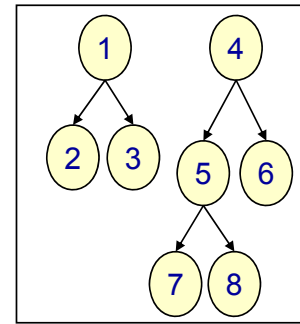
γ) **Inverse PageRank**





## Selecting Seeds: ( $\gamma$ ) Inverse PageRank

- επειδή η εμπιστοσύνη διαχέεται από τις καλές σελίδες, είναι λογικό να επιλέξουμε εκείνες τις σελίδες από τις οποίες μπορούμε να φτάσουμε σε πολλές άλλες
  - άρα μια ιδέα είναι να επιλέξουμε τις σελίδες με πολλά outlinks
  - Επιλογή των p1, p4, p5
- γενίκευση: επιλέγουμε τις σελίδες που δείχνουν σε πολλές σελίδες οι οποίες με τη σειρά τους δείχνουν σε πολλές σελίδες, **κ.ο.κ**
  - Επιλογή της p4
- Τρόπος: Αφού η σπουδαιότητα μιας σελίδας εξαρτάται από τα outlinks της (και όχι από τα inlinks της), μπορούμε να χρησιμοποιήσουμε την PageRank αντιστρέφοντας την φορά των ακμών



## Πειραματική Αξιολόγηση Experimental Evaluation



## Experimental Evaluation

- Experiments on the complete set of pages crawled and indexed by AltaVista (Aug. 2003)
- To reduce computational cost: work at the level of web sites (instead of web pages)
  - grouping of the (billions of) pages into 31 millions sites
  - websiteA points to websiteB if one or more pages from websiteA point to one or more pages of websiteB
    - So at most 1 link may start from website A and point to website B
  - Observations
    - 1/3 of the websites are unreferenced
    - So TrustRank **cannot differentiate** between them because they all have  $|\text{in}(p)|=0$
    - However they are low scored anyway (e.g. by PageRank) so they do not appear high in answers



## Experimental Evaluation: Seed Selection

- (1)  $\mathbf{s} = \text{SelectSeed}()$
- (2)  $\sigma = \text{Rank}(\{1, \dots, N\}, \mathbf{s})$
- (3)  $\mathbf{d} = \mathbf{0}_N$
- (4) for  $i=1$  to  $L$  do
  - if  $O(\sigma(i)) = 1$  then  $\mathbf{d}(\sigma(i)) = 1$

### Seed Set Selection

- Inverse PageRank applied on the graph of websites worked better than High PageRank (for the seed selection process)
- Parameters:  $a:0.85$ , iterations:20
  - With 20 iterations the relative ordering stabilized
- **Manual inspection of the top 1250 sites (i.e.  $|\mathbf{S}|=1250$ )**
- **From these only 178 were used as good seeds, i.e.  $|\mathbf{S}^+| = 178$  sites**



## Experimental Evaluation: Evaluation Sample

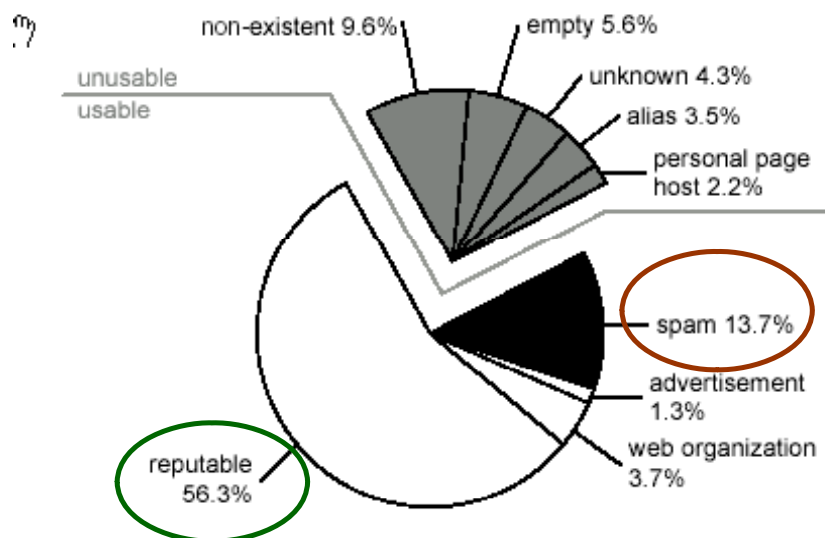
- To test the effectiveness of TrustRank we need a Reference Collection (e.g. something like TREC)
- A sample set X of 1000 sites was selected and evaluated manually, i.e. the oracle function was invoked (i.e. a person inspected them and decided whether they are spam or not)
- The Sample set X was not selected at random.
  - Recall that we are mainly interested in spam pages that appear high in answers
  - The following sample selection method was followed:

- Generate list of sites in decreasing order of their PageRank scores
- Segment them into 20 buckets so that the sum of the scores in each bucket equals 5% of the total PageRank score
  - |bucket1|=86, |bucket2|= 665, ... , |bucket20|= 5 millions pages
- select 50 sites at random from each bucket ( $20 * 50 = 1000$ )



## Experimental Evaluation: Evaluation Sample

The results of the manual evaluation (oracle invocation) of the pages in the sample set of 1000 sites:





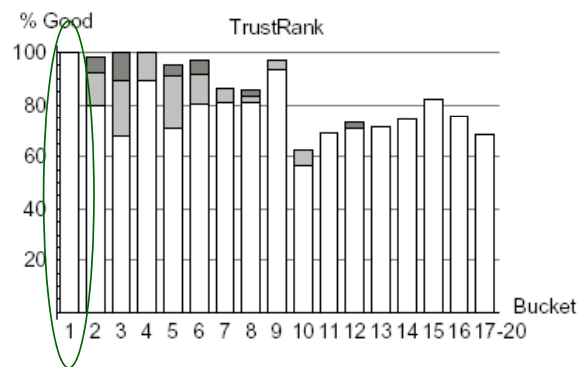
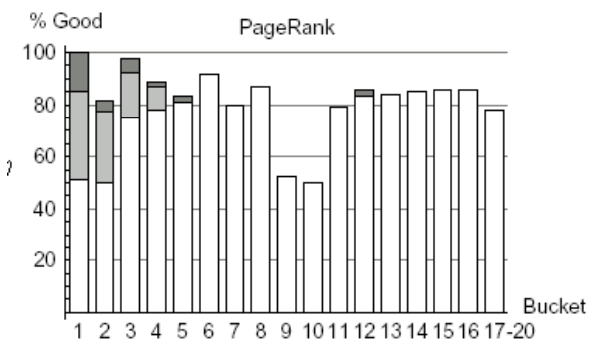


This collection (i.e. the set X) was used for evaluating TrustRank versus PageRank



## Evaluation Results: Comparing PageRank with TrustRank

### Good sites



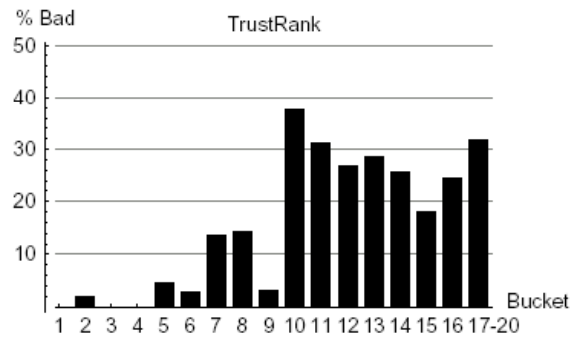
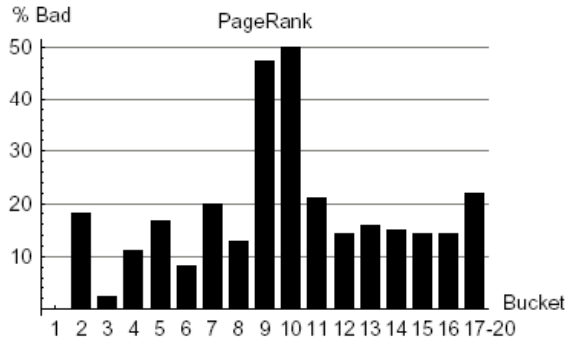
Reputable=white, advertisement =gray, webOrganization = dark gray

Notice that according to TrustRank the first bucket (that comprises 86 sites) has only reputable sites.



## Evaluation Results: Comparing PageRank with TrustRank

### Bad sites

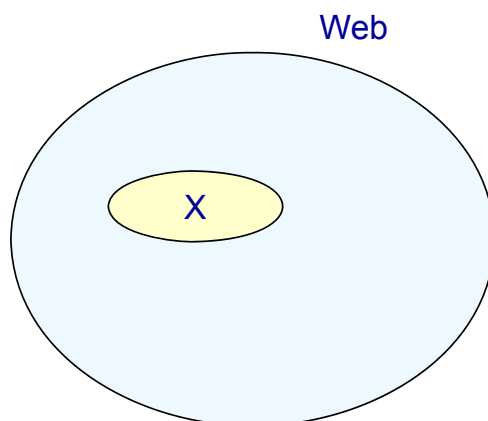


*TrustRank is a reasonable spam detection tool*



## Μέτρα Αξιολόγησης της Συνάρτησης Εμπιστοσύνης (Evaluation Metrics for the Trust Function)

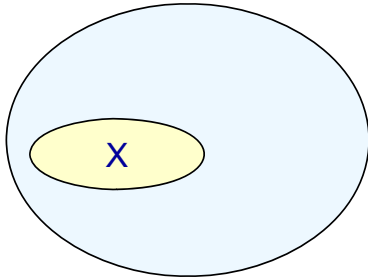
- Assume a **sample set X** of web pages for which we can invoke both T and O





## Μέτρα Αξιολόγησης της Συνάρτησης Εμπιστοσύνης: Precision and Recall

- We could evaluate a trust function  $T$  on the basis of an oracle function  $O$ .
- To this end we need to define appropriate measures.
- We can define two measures called precision and recall (analogous to the classical measures of Lecture 2) based on the **threshold** trust property:

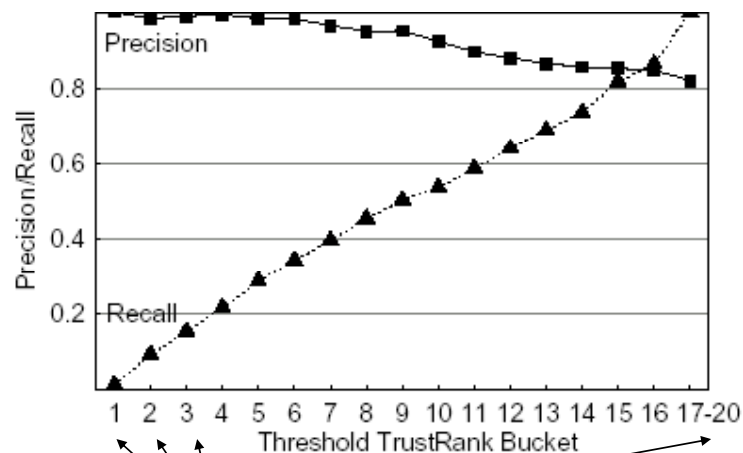


$$prec(T, O) = \frac{|\{p \in X \mid T(p) > \delta \text{ and } O(p) = 1\}|}{|\{q \in X \mid T(q) > \delta\}|}$$

$$rec(T, O) = \frac{|\{p \in X \mid T(p) > \delta \text{ and } O(p) = 1\}|}{|\{q \in X \mid O(q) = 1\}|}$$



## Μέτρα Αξιολόγησης της Συνάρτησης Εμπιστοσύνης: Precision & Recall: Πειραματική Αξιολόγηση

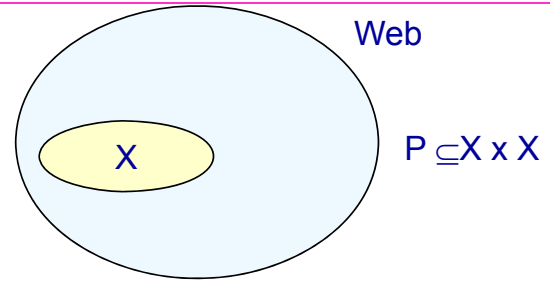


$\delta$ : such that to separate buckets



## Μέτρα Αξιολόγησης της Συνάρτησης Εμπιστοσύνης: Pairwise Orderedness

- We can generate from the set  $X$  a set  $P$  of pairs and we can compute the fraction of the pairs for which  $T$  did not make a mistake.



- The following metric can signal a violation of the **ordered trust property**

$$I(T, O, p, q) = \begin{cases} 1 & \text{if } T(p) \geq T(q) \text{ and } O(p) < O(q) \\ 1 & \text{if } T(p) \leq T(q) \text{ and } O(p) > O(q) \\ 0 & \text{otherwise} \end{cases}$$

$$\text{pairord}(T, O, P) = \frac{|P| - \sum_{(p,q) \in P} I(T, O, p, q)}{|P|}$$

- $\text{Pairord}(T, O, P) = 1$  if  $T$  does not make any mistake
- $\text{Pairord}(T, O, P) = 0$  if  $T$  makes always mistakes



## Μέτρα Αξιολόγησης της Συνάρτησης Εμπιστοσύνης: Pairwise Orderedness: Πειραματική Αξιολόγηση

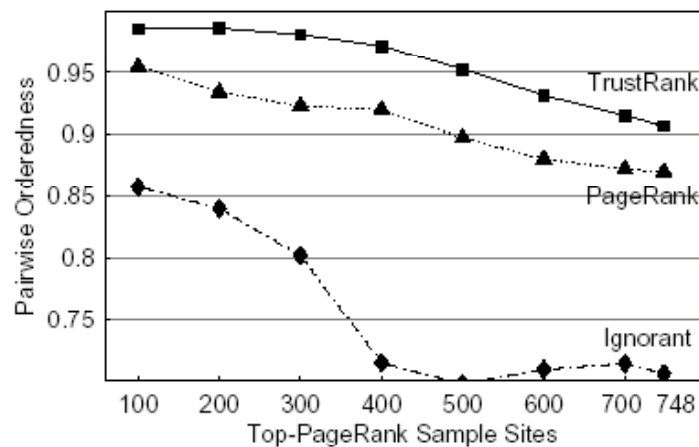


Figure 12: Pairwise orderedness.



TrustRank can effectively filter out spam from a significant fraction of the Web, based on a good seed set of less than 200 sites

### Σύνοψη

- Με αντιστροφή της φοράς των συνδέσμων και εφαρμογή του PageRank μπορούμε να προσδιορίζουμε ένα σύνολο που αξίζει αξιολόγηση από άνθρωπο
  - στην ουσία διαβαθμίζουμε τις σελίδες ως προς την αξία τους για χρήση στο δείγμα χειρονακτικής αξιολόγησης
- Ο αξιολογητής αποφαίνεται για την ποιότητα της κάθε σελίδας του δείγματος.
- Εκμεταλλευόμαστε την παραπάνω αξιολόγηση μέσω του TrustRank ο οποίος είναι στην ουσία ένας Biased Page Rank
  - η πιθανότητα τυχαίων αλμάτων είναι μεγαλύτερη προς τις καλές σελίδες



- Z. Gyongyi, H. Garcia-Molina, J. Pedersen, Compating Web Spam with Trust Rank, SIGMOD'04
- See also
  - Zoltan Gyongyi, Hector Garcia-Molina, Web Spam Taxonomy (<http://airweb.cse.lehigh.edu/2005/gyongyi.pdf>)