Topology-Based Spam Avoidance in Large-Scale Web Crawls

Clint Sparkman

Joint work with Hsin-Tsang Lee and Dmitri Loguinov

Internet Research Lab
Department of Computer Science and Engineering
Texas A&M University

Agenda

- Introduction
- Dataset and Ranking Algorithms
- Manual Analysis
- Automated Analysis
- Supporters Estimation
- Conclusions

Introduction

- Competition for high placement in search results has led to unethical Internet practices designed to deceive (spam) search engines in order to manipulate their ranking
- Web spam not only adversely impacts the quality of search results, but also impedes web exploration for a variety of research purposes
- Web crawlers must detect and avoid "undesirable" content in real-time

IRLbot

- High performance web crawler developed at the TAMU Internet Research Lab
- Able to perform several billion page web crawls with a single server
- Prioritizes queued pages using real-time snapshots of the Pay-Level Domain (PLD) graph

Pay-Level Domains

- PLDs must be purchased/acquired at a TLD or cc-TLD registrar
- PLD graphs offer some inherent advantages over other structures such as page-level or host-level graphs
  - More difficult and costly to manipulate, since PLDs must be registered, compared to links or hosts that can be trivially generated with scripts
  - Dramatically smaller graph that requires less processing and enables more efficient ranking during large crawls

Prioritization

- Crawlers need methods to budget their finite resources to spend most of their time exploring valuable parts of the Internet
- Prioritized web crawlers should be able to differentiate between domains that should be massively crawled and those that should not
- Two performance measures in achieving this classification
  - Accuracy: ability to avoid over-allocating resources to low-quality domains
  - Overhead: amount of processing required

Agenda

- Introduction
- Dataset and Ranking Algorithms
- Manual Analysis
- Automated Analysis
- Supporters Estimation
- Conclusions

Dataset

- IRLbot web crawl collected from June-Aug 2007
- Successfully downloaded 6.3B 200-OK HTML pages
- Webgraph has 41B nodes and 310B edges
- Host graph has 641M nodes and 6.8B edges
- PLD graph has 89M nodes and 1.8B edges

Ranking Algorithms

- In-degree (IN) – Sum of in-links
- Supporters (SUPP) – Let $d(i,j)$ be the shortest distance from i to j along the directed graph G

$$SUPP(j) = \sum_{i=1}^{n} 1_{d(i,j)=2}$$

- PageRank – Models a random walker on G, where the walker traverses an out-link with probability $\alpha = 0.85$ or teleports to a random node with probability $1 - \alpha$
- Weighted In-degree (WIN) - $WIN(j) = \sum_{i:(i,j)\in E} \frac{1}{d_{out}(i)}$

Agenda

- Introduction
- Dataset and Ranking Algorithms
- Manual Analysis
- Automated Analysis
- Supporters Estimation
- Conclusions

Manual Spam Evaluation

- There is no common algorithm to measure ranking results in spam avoidance applications
- Previous work manually classified a small random sample of the graph as good/bad. Competing rankings are divided into K buckets and compared based on the buckets where the spam is found
- Our approach: manually scrutinize the top-1K PLDs in each prioritized ranking

Manual Spam Evaluation, cont.

- No consensus on the definition of spam
  - Is pornography spam?
- We use a *subjective* approach using the following criteria
  - Attempts to perform malicious activities upon visit (malware or virus)
  - Overwhelming presence of links whose primary purpose is to generate revenue from click-throughs
  - No immediate useful content can be discerned in the PLD

Google Toolbar Rank (GTR)

- Google offers a toolbar for web browsers that, among other things, offers a quantitative value from 0-10
- Some pages have no GTR
  - Page has not been crawled
  - No longer exists
  - Purposefully removed from the index
- No ranking analysis has previously involved GTR values

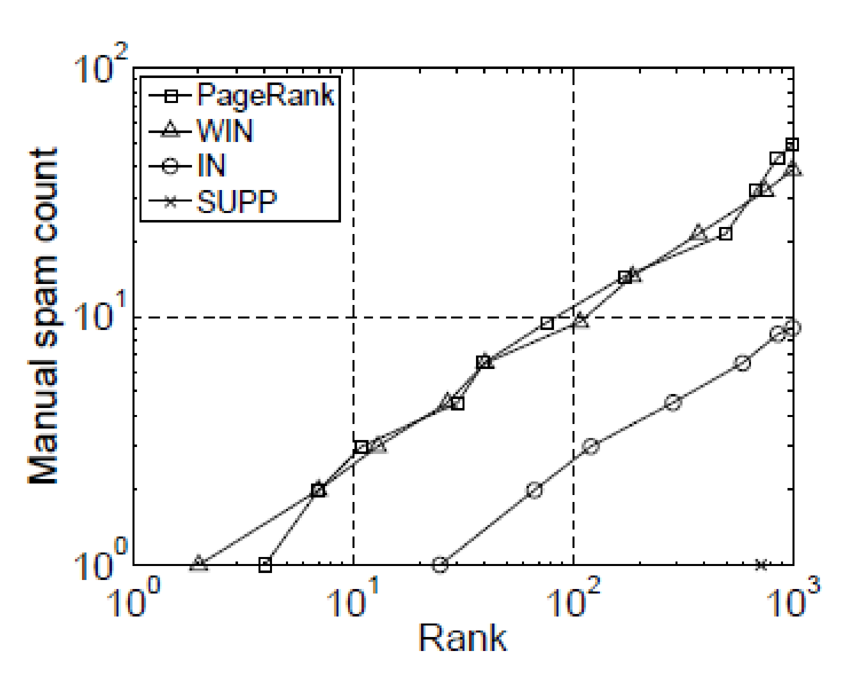
Top-ranked PLDs

IN PLD	GTR	PageRank PLD	GTR	WIN PLD	GTR	$SUPP_2$ PLD	GTR
microsoft.com	9	microsoft.com	9	microsoft.com	9	google.com	10
google.com	10	adobe.com	10	information.com (S)	5	microsoft.com	9
yahoo.com	9	google.com	10	google.com	10	yahoo.com	9
adobe.com	10	information.com (S)	5	adobe.com	10	adobe.com	10
blogspot.com	9	macromedia.com	10	macromedia.com	10	macromedia.com	10
wikipedia.org	9	yahoo.com	9	yahoo.com	9	wikipedia.org	9
w3.org	10	sedoparking.com (S)	–	sedoparking.com (S)	–	blogspot.com	9
geocities.com	9	googlesyndication.com	–	miibeian.gov.cn	9	msn.com	8
msn.com	8	w3.org	10	googlesyndication.com	–	apple.com	9
amazon.com	9	miibeian.gov.cn	9	w3.org	10	geocities.com	9
aol.com	8	downloadrings.com (S)	1	ndparking.de (Q)	–	w3.org	10
myspace.com	9	chestertonholdings.com (Q)	–	statcounter.com	9	sourceforge.net	9
macromedia.com	10	juccoholdings.com (Q)	–	searchnut.com (S)	–	youtube.com	9
youtube.com	9	statcounter.com	9	revenuedirect.com (Q)	4	bbc.co.uk	9
tripod.com	7	linkz.com (Q)	3	myspace.com	9	netscape.com	8

- IN and SUPP PLDs are much more reputable (large, well-known domains) and contain no spam
- PageRank and WIN promote spam/questionable domains to the top of their ranked lists

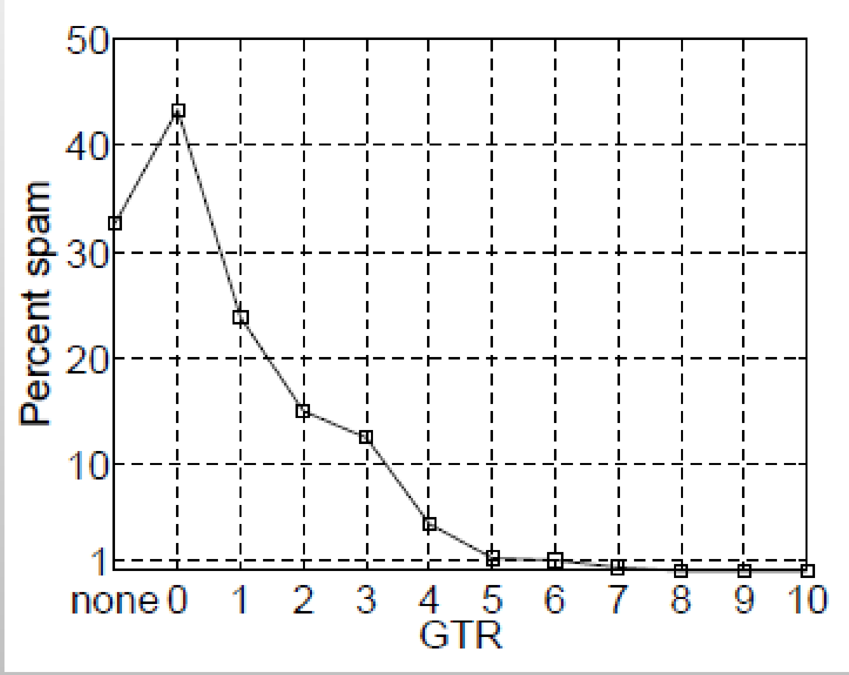
Spam Avoidance

- Compare the amount of spam found in the top-1K for each
- PageRank and WIN similar performance - 49 and 39 spam sited in top-1K
- IN allows 9 in the top-1K, the first in pos 25
- SUPP allows only 1, linksynergy.com in pos 718



GTR and Spam

- Examine how well GTR predicts spam
- 2,100 PLDs manually examined (aggregate of all top-1K lists)
- No GTR-0 sites were well-known, reputable sites
- Almost no spam (0.6%) occurs at GTR 5 or higher

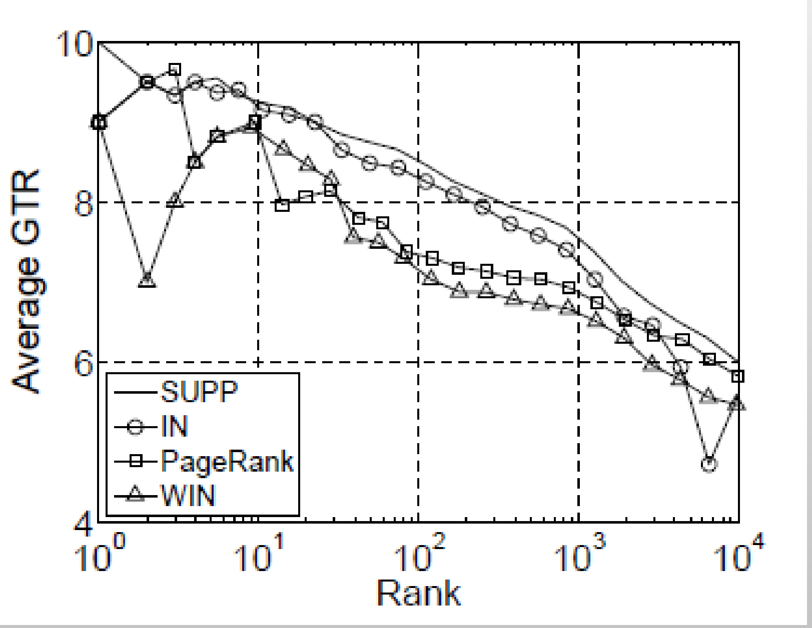


Agenda

- Introduction
- Dataset and Ranking Algorithms
- Manual Analysis
- Automated Analysis
- Supporters Estimation
- Conclusions

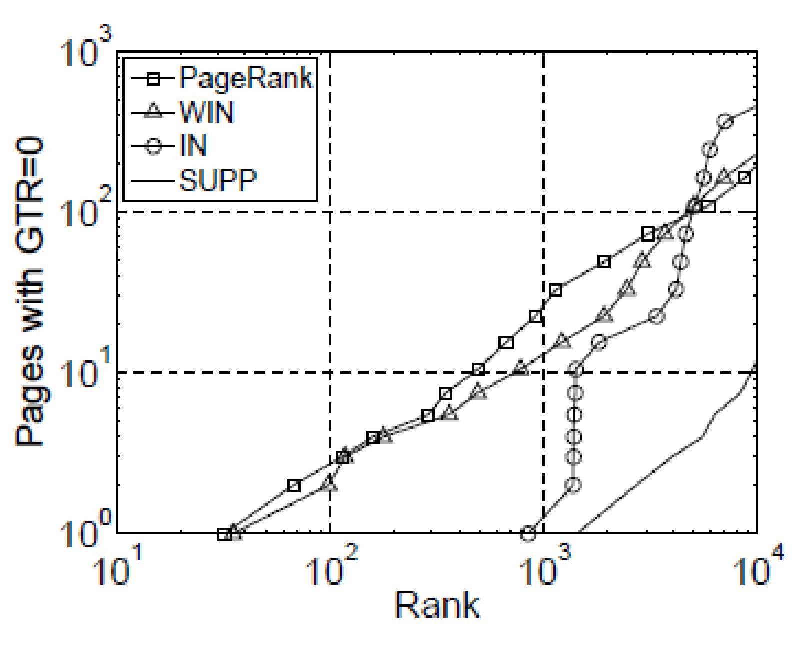
Average GTR

- Graph plots a running average of the GTRs
- Compares how well each algorithm places the most valuable PLDs at the top of the list
- In has a sharp drop after 4K. Many GTR-0 PLDs related to worldnews.com



GTR-0

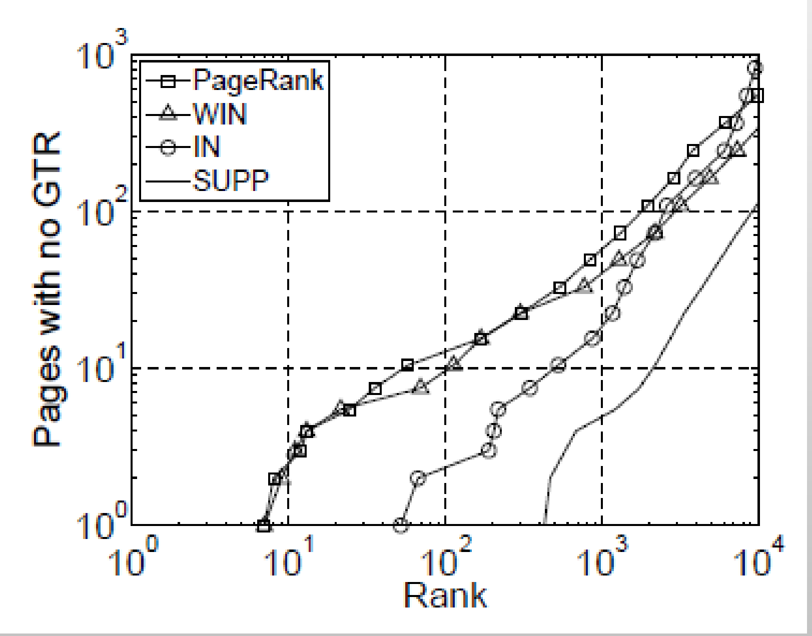
- Cumulative distribution of PLDs with GTR 0
- SUPP does not allow a GTR-0 PLD until pos 1,422
- IN initially does very well. Only 1 in top 1K, pos 843, but worldnews.com sites quickly add around 2K
- Both PageRank and WIN allow GTR-0 PLDs high in their rankings (32 and 35)



Computer Science, Texas A&M University

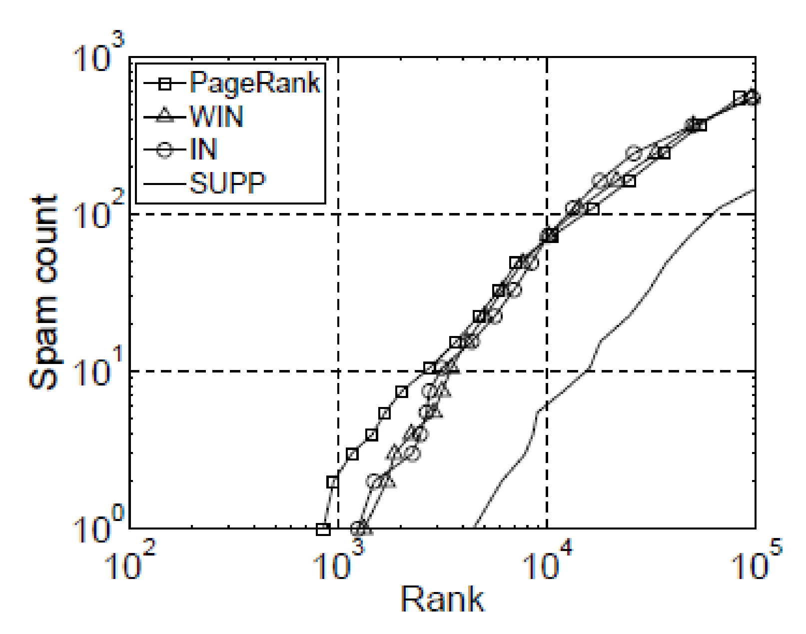
No GTR

- Cumulative distribution of PLDs with no GTR
- SUPP is the clear winner with 1st in pos 469
- PageRank and WIN are very similar. Both allow 4 PLDs with no GTR in their top-15
- IN only allows 2 in top-100, but has the most in top-10K



Blacklisted PLDs

- Cumulative distribution of PLDs on SpamAssassin's blacklist and considered to be related to email spam
- SUPP is the clear winner, with the 1st PLD in pos 4,459, and only 7 in the top-10K



High GTR

- To understand if any good domains ended up in the bottom of our lists, we examine the 470 PLDs with GTR 9 or 10 that appear in SUPP’s ranking past 10K
- All fall within the following 4 categories
  - Redirects to famous domains for either misspelled or unknown domains, or country versions of main site
  - Mirrors that do not redirect to main site, but look identical
  - .gov or .edu sites that Google commonly inflates
  - GTR anomalies that have been since corrected

Depth of Supporters

- Next explore if SUPP at depth 2 is the best choice for Internet graphs
- We found SUPP at depth 3 to be a poor indication of PLD reputation
  - Due to the rapid explosion of supporters for popular PLDs and the lack of nodes to reach at depth 3
  - google.com
    - Highly ranked by all algorithms
    - 15.5M level-2 supporters vs 6.2M level-3 supporters
  - hotsitekey.info
    - Ranked on position 192,056 by $SUPP_2$
    - Manages 15.6M level-3 supporters

Agenda

- Introduction
- Dataset and Ranking Algorithms
- Manual Analysis
- Automated Analysis
- Supporters Estimation
- Conclusions

Estimating Supporters

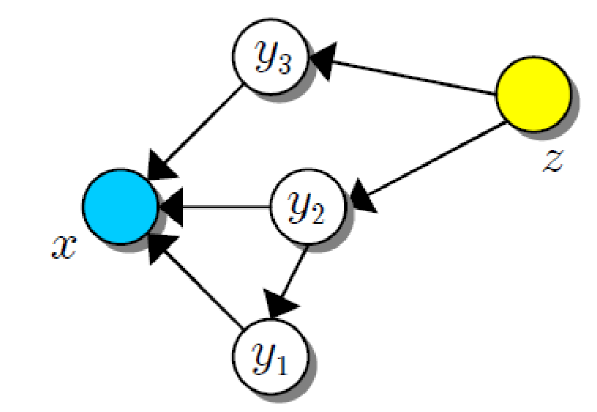
- Easy to see that SUPP produces the best ranked PLD lists
- Calculating SUPP directly does not scale well to large graphs due to the enormous amount of processing to perform a limited BFS search from each node
- Good news: a high-performance crawler only requires a *fast, accurate, and scalable* technique for estimating supporters at the top of the list

Quick Visit Supporters

- Quick Visit Supporters (QVS) simply counts the number of link traversals during BFS

$$QVS(j) = \sum_{i:(i,j)\in E} d_{in}(i)$$

- High error due to duplicate node counts



(a) SUPP_2

Bit Vector Estimate

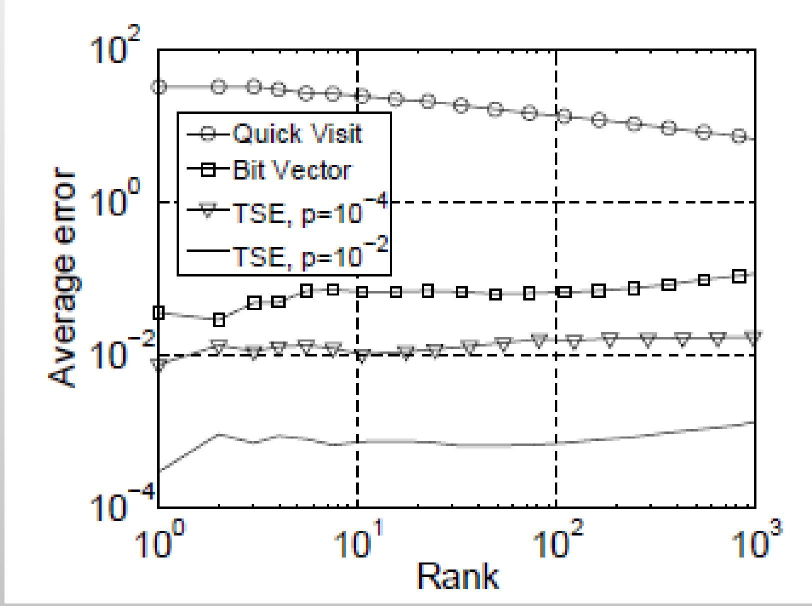
- Nodes iteratively receive bit strings from their neighbors and apply a bitwise OR against their own bit string
- Length of bit vector determines accuracy.
  - 64-bit vectors used in comparison
- Requires $log_2(S_{max})$ rounds to terminate, where S_{max} is the maximum SUPP count
  - 25 rounds for IRLbot's PLD graph
  - Can adapt to only 2 rounds if we only need to estimate the top-1K or 3 rounds for top-10K

Top Supporters Estimate (TSE)

- Scan the out-graph and retain in RAM a p-fraction of all nodes z along with their adjacency lists $\{w_j\}$
  - Produces an unbiased random sample of all supporters z that x will later count
- Sequentially read the in-degree graph, and examine each node x with its neighbors $\{y_i\}$
  - If $x \neq z$ and $x \notin \{w_j\}$, any overlap between $\{y_i\}$ and $\{w_j\}$ indicates z is a supporter of x at level 2
- Scale the supporter count for x by $1/p$

Estimation Error

- Error is calculated against the true SUPP count
- Plot is by true SUPP rank
- Quick Visit has enormous error (> 1,000%) for the top PLDs
- Bit Vector error averages 6.5% in this range
- VSE error averages ~ 1% for $p = 10^{-4}$ to 0.1% for $p = 10^{-2}$



Comparison – RAM only

Algorithm	Hits	Ops	Time	Speedup
$SUPP_2$	4.9T	1.9T	70 hrs	–
TSE ($p = 10^{-2}$)	49B	19B	11 min	381
Bit Vector ($r = 2$)	7.1B	11B	3.8 min	1,113
TSE ($p = 10^{-3}$)	4.9B	1.9B	70 sec	3,600
Quick Visit	1.8B	1.8B	55 sec	4,581
TSE ($p = 10^{-4}$)	490M	190M	7.5 sec	33,600

- Table shows the theoretical number of random RAM hits and various CPU operations
- Time is actual running time on Quad-CPU server with enough RAM to hold entire PLD graph
- Speedup factor is compared to SUPP

External Memory Techniques

- SUPP-A: loads sequential chunks of the in-graph and then re-scans the entire in-graph
- SUPP-B: simultaneously reads in/out graphs and writes out all pairs (x, z) where z is x's level-2 supporter. A k-way merge is performed to eliminate duplicates.
- Quick Visit: reads the file twice and stores the last vectors of in-degree counts and hashes
- VSE: reads in/out graphs but does not require that all supporters counts fit in RAM

External Memory Comparison

Algorithm	Disk read	Disk write	RAM	Phases
SUPP_2-A	32 TB	–	8 GB	–
	130 TB	–	2 GB	–
	2.6 PB	–	100 MB	–
SUPP_2-B	49 TB	49 TB	8 GB	1
	98 TB	98 TB	2 GB	2
	147 TB	147 TB	100 MB	3
Bit Vector ($r = 2$)	63 GB	–	1.9 GB	–
Quick Visit	31.4 GB	–	2.1 GB	–
TSE ($p = 10^{-2}$)	31.4 GB	–	157 MB	–
TSE ($p = 10^{-3}$)	31.4 GB	–	16 MB	–
TSE ($p = 10^{-4}$)	31.4 GB	–	1.6 MB	–

- I/O complexity using 15.8GB PLD graph with 8-byte hashes
- SUPP-A (8GB RAM) reads the graph 2,000 times!
- SUPP-B scales better with reads, but requires an enormous amount of disk to write to
- TSE has constant I/O, and RAM is determined by p (accuracy)

Agenda

- Introduction
- Dataset and Ranking Algorithms
- Manual Analysis
- Automated Analysis
- Supporters Estimation
- Conclusions

Conclusions

- This paper compared various agnostic algorithms for ranking the web at the PLD level
- Leveraged manual analysis and Google Toolbar Rankings for automated analysis
- SUPP decisively outperformed the other techniques but was infeasible in practice
- Top Supporters Estimate is a fast, accurate, and scalable estimator for the top-ranked PLDs