

Countering Web Spam with Credibility-Based Link Analysis

James Caverlee

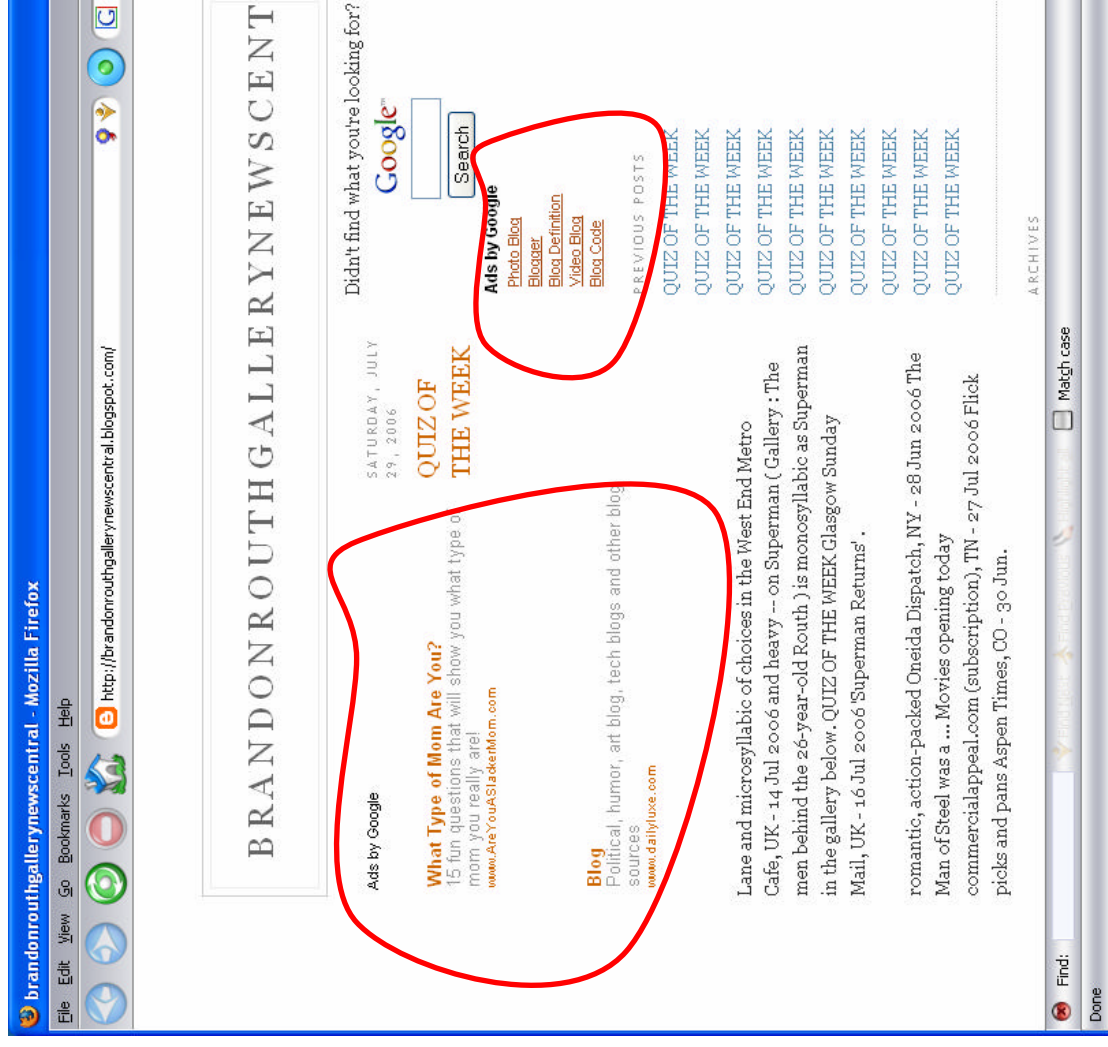
College of Computing
Georgia Institute of Technology



The Wild Wild Web

- Page Spoofing
 - **2m** users tricked, resulting in **\$1.2b** in losses for U.S. banks and card issuers in 2003 (Harvard)
- Spyware/Malware
 - Bundled with seemingly innocent software
 - Drive-by downloads
 - **17%** of US websites host spyware; **42%** of Chinese websites (Webroot)
- \$\$\$
 - Web ad revenue: **\$12b** in 2005
 - **9%** of new blogs are machine-generated (Technorati)

CERCS IAB April 27, 2007

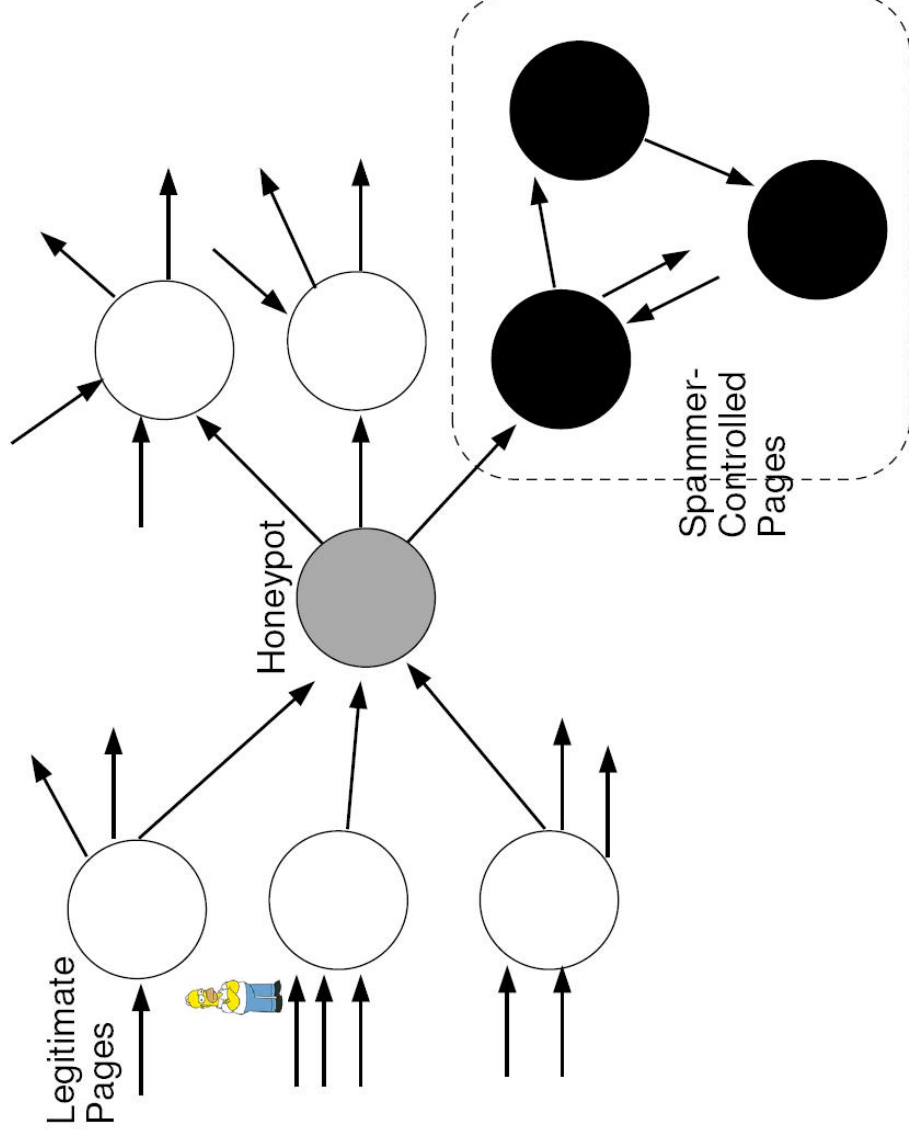


Strategies Used by Spammers to Trap Users

- Social engineering
- Email spam (*phishing*)
- DNS cache poisoning (*pharming*)
- Infiltrate Web directories (like Yahoo!)
- **Corrupt search engines** →

~8% of pages and ~18% of sites are designed to manipulate search engines [2005]

Vulnerabilities in Link-Based Search Systems: **Honeypots**

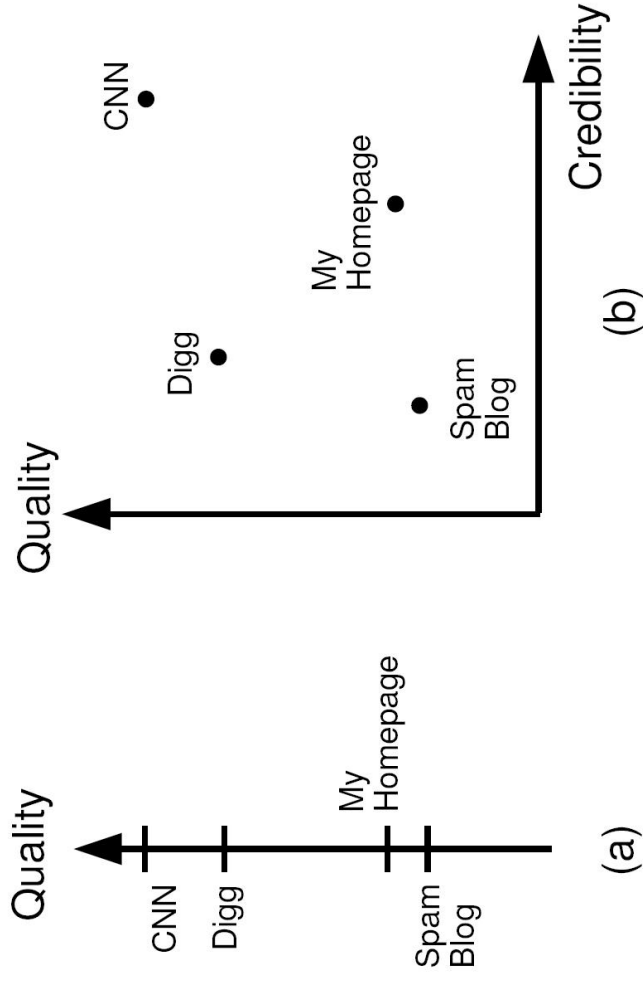


Combating Web Spam: Challenges & Approaches

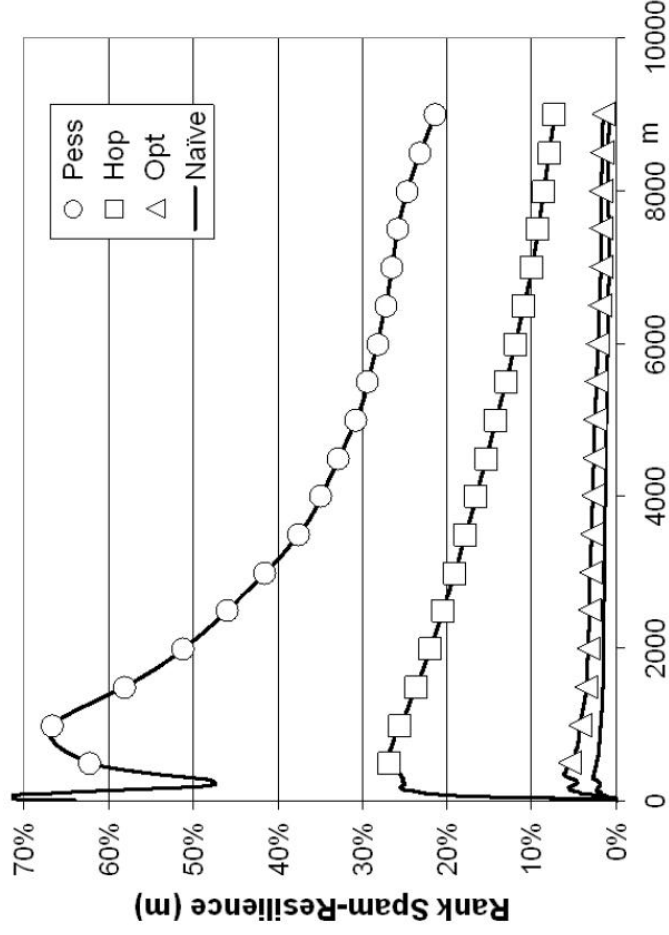
- Challenges
 - Large and growing Web
 - 11+ billion indexable Web pages (2005)
 - Low costs to spammers, High return (\$\$\$)
 - Smart, deceptive adversary
- Approaches
 - Gated Web
 - Spam detection and removal
 - **Spam-Resilient Algorithms**
 - **Source-centric ranking** [Caverlee, et.al: IPDPS 07, IEEE CollaborateCom 06]
 - **Credibility-based link analysis** [Caverlee & Liu: ACM PODC 07]

Credibility-Based Link Analysis

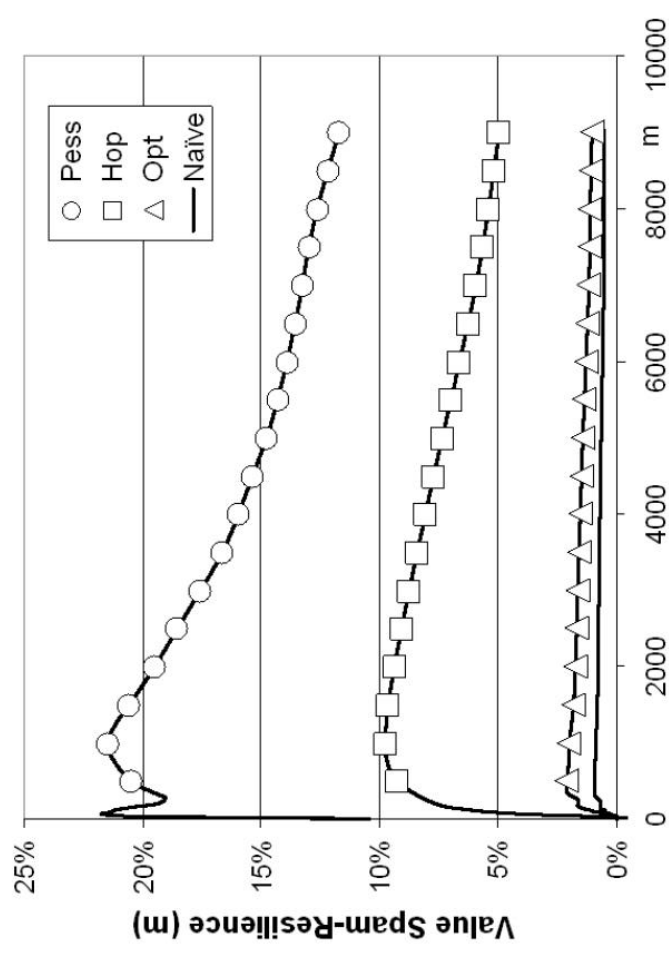
- Link-based attacks
 - Corrupt the notion of a link as an “endorsement”
- Idea:
 - Distinguish the quality of a page’s links (**link credibility**) from the quality of the page



CredibleRank vs. TrustRank: Rank and Value Spam-Resilience



(a) $SR_{Rank}(m)$



(b) $SR_{Value}(m)$

Summary of Contributions

- Introduced **credibility-based link analysis**
- Tunable k-Scoped credibility
 - Explored key factors – k, discount factor, etc.
 - Completely **personalizable**
 - Depending on your definition of spam
 - Depending on your risk tolerance
- CredibleRank
 - **Strongly counters** link-based manipulation
- Experimental Results
 - High-quality results with small scope (k) and small blacklist
 - More spam-resilient than PageRank and TrustRank

Related Research & PhD Dissertation



- Web Spam Detection
 - Using Email Spam (CEAS '06)
- Spam-Resilient Search Architectures
 - P2P-Based (IEEE ICDE '07)
 - Quality-Aware Distributed IR (ACM SIGIR '06)
- Spam-Resilient Ranking
 - Influence Throttling (IEEE IPDPS '07)
 - Source-Centric (IEEE CollaborateCom '06)
- Other Related Work:
 - Web Categorization and Integration Services (ACM ICSOC'04, WWW'05, WWWJ'06)
 - Deep Web Search and Data Extraction (ICDE'04, TKDE'05, IJWGS, SCC'05)
 - Workflows and Process Mining (ICWS'06, BPM'06)

Thank you