# Microsoft Silicon Valley Web Spam Challenge Entry

Steve Chien, Dennis Fetterly,

Mark Manasse, Marc Najork, and Alexandros Ntoulas

# Our Approach

- Use machine learning to combine
  - Features provided by the organizers
  - Features from our previous work
  - New features where we focus on those that have some creation cost

# Additional Features

- URL
  - Features derived from the URLs in the UK 2006-05 dataset, such as the number of dots, dashes, and digits in the hostname
- Page Content
  - Features derived from word frequency analysis in the 77 million pages
  - Features based on grouping documents into sets of near-duplicate documents
- Graph Structure
  - Features indicative of link exchange based on the UK 2006-05 page-level and host-level web graph
- Evolution
  - Results from re-crawling the 77 million URLs in the UK2006-05 dataset
  - URL overlap with web crawl that occurred in 2002
- Economic
  - Features derived from the registrar records for the 7,707 domains in the UK 2006-05 dataset
  - Features based on the publisher ID of any Google AdSense advertisements embedded in the 77 million pages

# Feature Selection

- Ended up with 322 features
- Avoid over-fitting by selecting only most discriminating features
- Used feature selection algorithms in WEKA
  - Evaluated features using several attribute evaluators, search methods, and cutoff values
- The 75 features used were identified using the Ranker search method and the ChiSquared attribute evaluator.

# Top 10 Ranked Features

Observations:

- Validates that TrustRank (6/10) and neighborhood are important (2/10) for spam detection
- Economic features are also important

| Rank | Feature |
|---|---|
| 1761.73 | average_spamicity_neighbors_PASS2 |
| 1756.16 | log_OP_trustrank_hp_div_pagerank_hp_CP_ |
| 1741.12 | log_OP_trustrank_hp_div_indegree_hp_CP_ |
| 1518.52 | average_spamicity_neighbors_PASS1 |
| 1283.88 | L_trustrank_hp |
| 1177.37 | log_OP_trustrank_mp_div_pagerank_mp_CP_ |
| 1131.78 | MeanHostsAdSenseId |
| 1128.39 | log_OP_trustrank_mp_div_indegree_mp_CP_ |
| 1032.10 | L_trustrank_mp |
| 838.15 | STD_83 |

# Evaluation

- Evaluation on the 5,622 labeled hosts
- Used ten-fold cross validation
- Best classifier used bagging in combination with a C4.5 decision tree

| Class | Recall | Precision |
|---|---|---|
| Non-spam | 96.8% | 97.6% |
| Spam | 70.5% | 80.2% |

# Acknowledgements


- Challenge Organizers
- UK2006-05 collection coordinators