Extracting Link Spam using Biased Random Walks from Seed Sets

Baoning Wu (SNAP) and Kumar Chellapilla (Microsoft Live Labs)
Web Spam

• Actions intended to mislead search engines into ranking some pages higher than they deserve
  – Relevance (to query) and Importance (global popularity)

• Effects
  – Drop in quality of results
  – Inflated indices (server cost)
Link Spam

• Deliberately manipulate hyperlinks between web pages to unduly boost search engine ranking
  – One of the most popular search engine spamming techniques
  – Target link based ranking algorithms such as HITS, PageRank, etc

• Link Farms and Link Exchanges
Link Spam

• Deliberately manipulate hyperlinks between web pages to unduly boost search engine ranking
  – One of the most popular search engine spamming techniques
  – Target link based ranking algorithms such as HITS, PageRank, etc

• Link Farms and Link Exchanges
Two Types of Link Spam

• Link exchanges
  – Practice of exchanging links between websites. One or both parties involved solicit the other for reciprocal links.

• Link farms
  – A cluster of densely interconnected web sites or pages
  – Few owners, duplicated/artificial/scraped content
Extracting Link Spam

• Detection vs Extraction

• Extraction is a directed approach
  – Starts with a small seed set (marked spam) and simulates a local random walk to extract the spam community around the seed set

• Can be used interactively
  – Seed set is manually labeled
  – Groups of nodes in the extracted spam community are interactively evaluated

• Produces higher precision than fully automated methods
Link farm example

Here is the EasyWay™ Link Block that RunAway GetAway Alliance members proudly post on their destination vacation rentals websites:

Member of the RunAway GetAway Vacation Rentals Alliance - ©2005 Keligo®
Click Destination Name for Member Website, Click State Abbreviation for Alliance Destination Page
Santa Fe NM | North Lake Tahoe CA | Pagosa Springs CO | Sedona AZ | Coeur d'Alene ID | Hood River OR | British Columbia BC | Flagstaff AZ | Traverse City MI | Teton NV | Aspen Snowmass CO | Park City UT | Redwood NM | Cambria CA | South Lake Tahoe CA | Mammoth CA | Coastal Rhode Island RI | Boothbay Harbor ME | Monterey Bay CA | Cape Cod MA | White Mountains NH | Trinidad CA | Bainbridge Island WA | Steamboat Springs CO | Yellowstone Park MT | Maui HI | Newport Beach CA | Maui UT | Granada CA | South Coast Maine ME | Breckenridge Copper Mtn CO | Crested Butte CO | Kauai HI | Keystone CO | Telluride CO | Kehalani Coast HI | Outer Banks NC | Martha Beach SC | Adirondack Mountain NY | Lake George NY | Napa Sonoma CA | Bar Harbor ME | San Juan Islands WA
Sponsors: Keligo | TravelArb | New MediaWorks

WHY RUNAWAY GETAWAY?

Any home owner can list their vacation rentals with a listing service for a fee regardless of the quality or suitability of the property because listing services do not care. Such services are simply directory-driven.

RunAway GetAway is a national alliance for independent vacation rentals. We offer a list of locale-specific managers of vacation rentals -- Alliance Members -- across the county representing superior privately owned getaways and lodgings within their exciting vacation destinations. Each member organization is independently owned and operated. We also provide a dedicated page for each destination with links to points of interest, weather, maps, and other useful information. Click on a destination below to go to the Destination Page, for lodging information and much more.

Santa Fe | Mammoth | Breckenridge Copper Mtn
Adjacency matrix for link farm
Link exchange web page

Add a link to my website on yours, and I will link you here also! It's easy! Follow the instructions below. Once my link has been added to your site, just complete the form at the bottom of the page and I will then add your link to my site. Please add the following link to your Links page. You can simply copy & paste the text onto your site.

**YourNewHomeToday.com - Rent to own homes and rental houses in Louisville Kentucky and surrounding areas** - Rent to own homes, rental houses, and lease option homes in the Louisville Kentucky area. (Real Estate agents available for help).

**Title:** YourNewHomeToday - Rent to own homes and rental houses in Louisville Kentucky and surrounding areas  
**URL:** http://www.YourNewHomeToday.com  
**Description:** Rent to own homes, rental houses, and lease option homes in the Louisville Kentucky area. (Real Estate agents available for help).

Submit Info for Link Exchange

Name of your site: [Enter name]
Link exchange

B

A

C

D

E

leaf

HUB
Link exchange tendency
Extracting Link Spam

• Given one or more link spam sites, can we detect their partners?

• In other words, can we detect the community around the link spam seed sites?
Related work

• Spielman and Teng (STOC’2004)
  – Nearly-linear time algorithms for graph partitioning

• Andersen and Lang (WWW’2006)
  – Communities from seed sets

• Extract communities around seed sets based on spectral properties (conductance)
  – Correlates well with personalized/local page rank
(Lazy) Random walk

• We simulate the following random web surfer behavior

• Start from the seed node(s)
  – Set their probability to $1/|S|$

• At each step, from each node with non-zero probability
  – With 50% chance follow one of the out-links with equal probability
  – With 50% chance stay at the current node
    • (equivalent to jumping to another non-zero node in proportion to their current probability value)
Transition Probability Matrix

• Transition probability matrix at step t

\[ p^{t+1} = \frac{1}{2} \left( I + AD^{-1} \right) p^t \]

• n = number of nodes in the graph G
• P = probability vector (n x 1) over nodes in G
• A = adjacency matrix (n x n) of G
• D = diagonal matrix (n x n) where D(i,i) = degree(node i)
Resource redistribution
Resource redistribution

Spam seed
Resource redistribution

Initialization
Resource redistribution

First iteration
Resource redistribution

Second iteration
Biased Random Walk

• Bias the random walk to be local
  – Truncation (threshold, percentile)
    • Set very small probabilities to zero
  – Decayed random walk
    • Contain random walk in the neighborhood of the seed set
    • Exponential drop with distance from the seeds
  – White lists and black lists
    • Known good and bad sites
      \[ p^t[i] = p^t[i] \times \gamma[i] \]
    • Renormalization
      \[ \gamma[i] = 2^{-\delta(i)} \]
  • Compensate for leaked probability

• The random walk converges after a few iterations
Decayed version

A \rightarrow B
B \rightarrow C
C \rightarrow F
D \rightarrow B

w1 = 2w2
w1 = w2 = w2

Decayed version
Experiments

• Data set
  – Domain graph of July, 2006 (Live Search)
    • about 48M nodes and 470M directed edges
  – Seeds
    • Manually labeled 75 link farms
      – 46 big link farms (bigger than 50 members)
      – 27 small link farms (less than 50 members)
      – 2 blog link farms
    • Manually labeled 50 link exchange hubs
  – White list (25K domains)

• Directed, Inverted, and Undirected random walks
  – Each seed is processed separately
  – Simulated till convergence (10-30 iterations - a few seconds)
Evaluating Spam Communities

• Sort all the output nodes using on their convergence probability
• Distribute them into ten buckets based on percentiles
  – Bucket 1 has the top 10 percent of the nodes with the highest value
• From each bucket, randomly choose three nodes for manual evaluation.
  – For link farm, evaluate whether it is a member
  – For link exchange, evaluate whether it is a hub
Link Farm Results

Spam community sizes
(mean = 268)

Precision

Extracted Link Farm Community Size

Mean Precision

Bucket
Link Exchange Results

Spam community sizes
(mean = 513)

Extracted Link Exchange Community Size

Precision

Mean Precision

Bucket

0 2 4 6 8 10