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

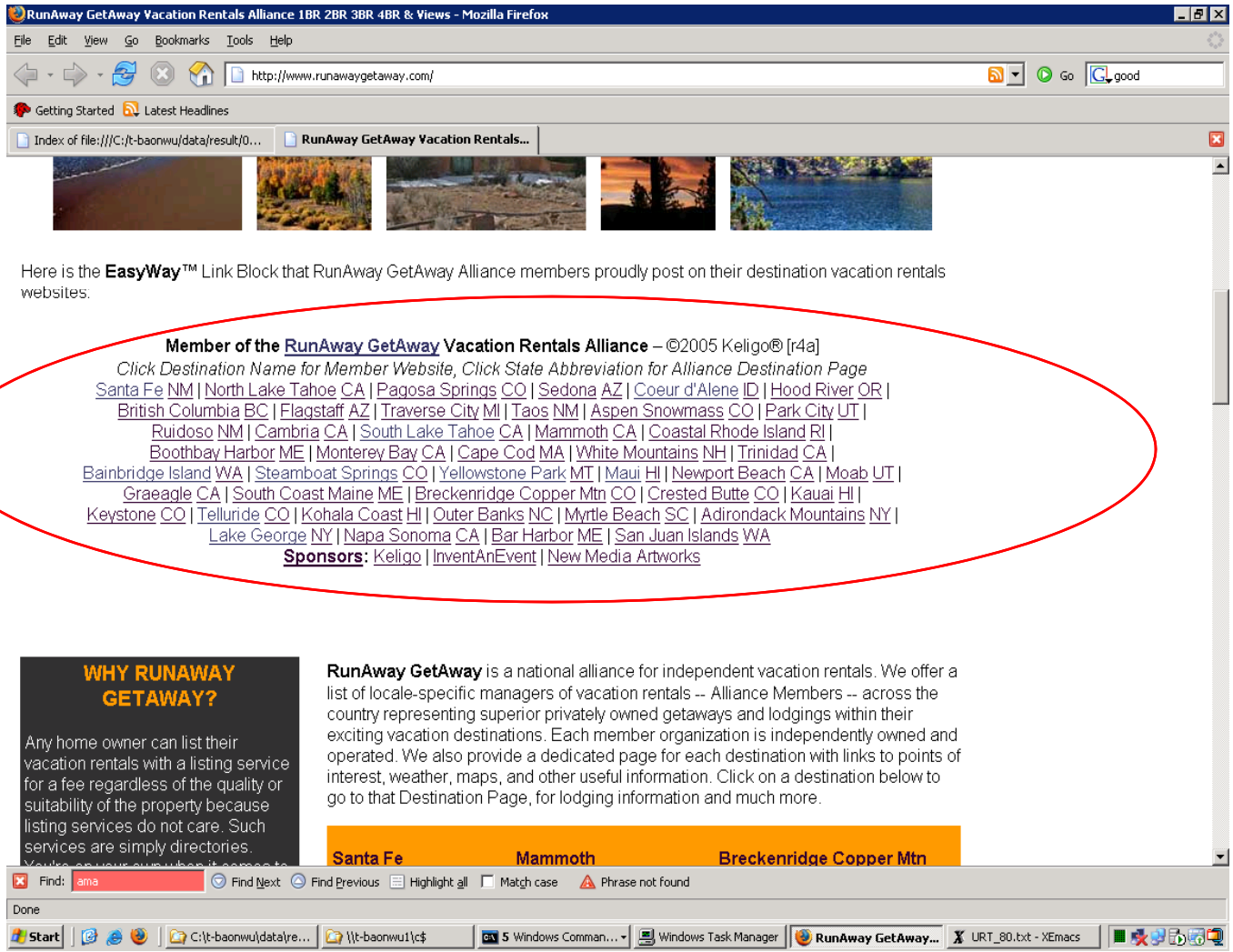
Two Types of Link Spam

- 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

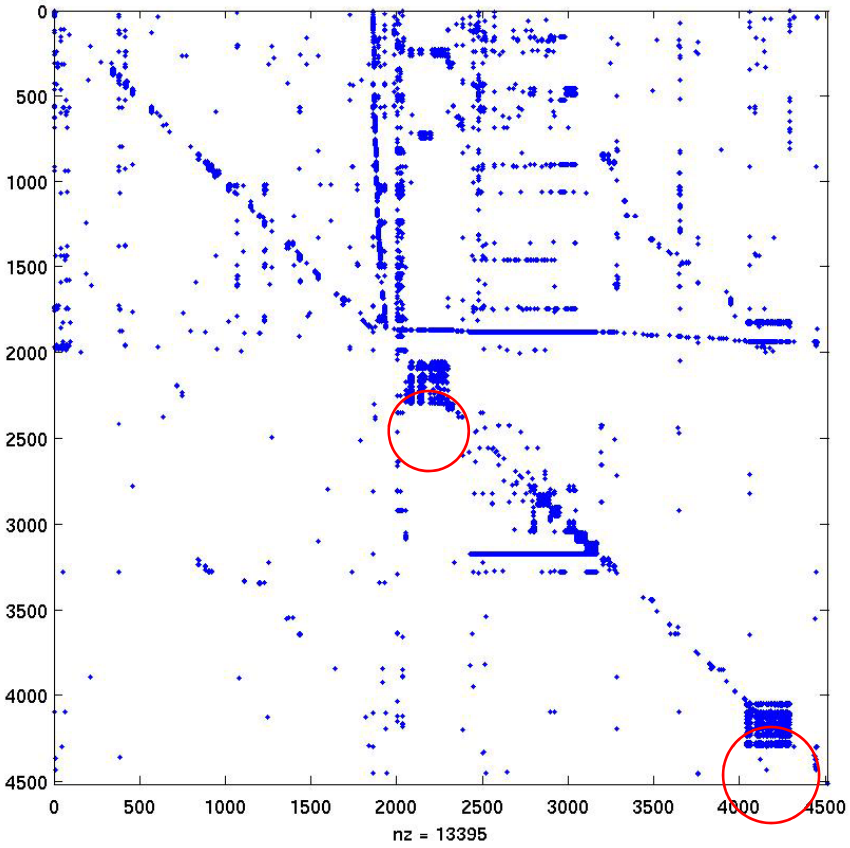
Extracting Link Spam

- 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

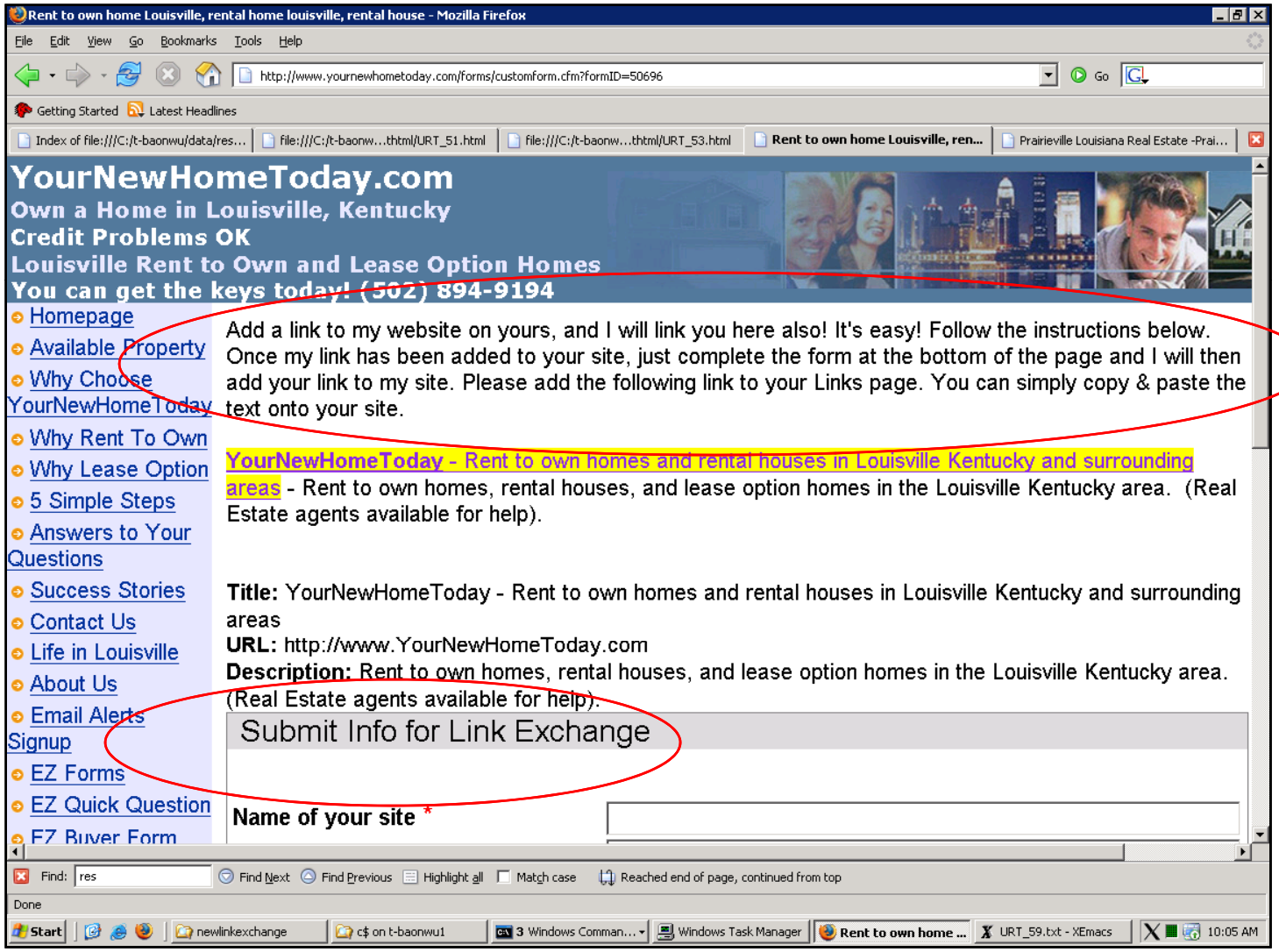
Link farm example



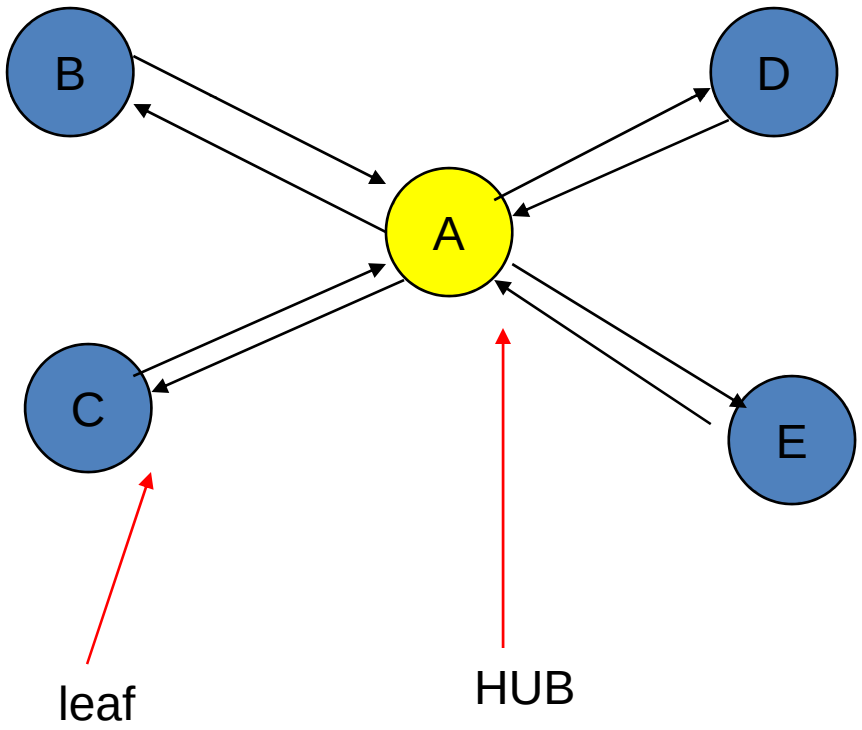
Adjacency matrix for link farm



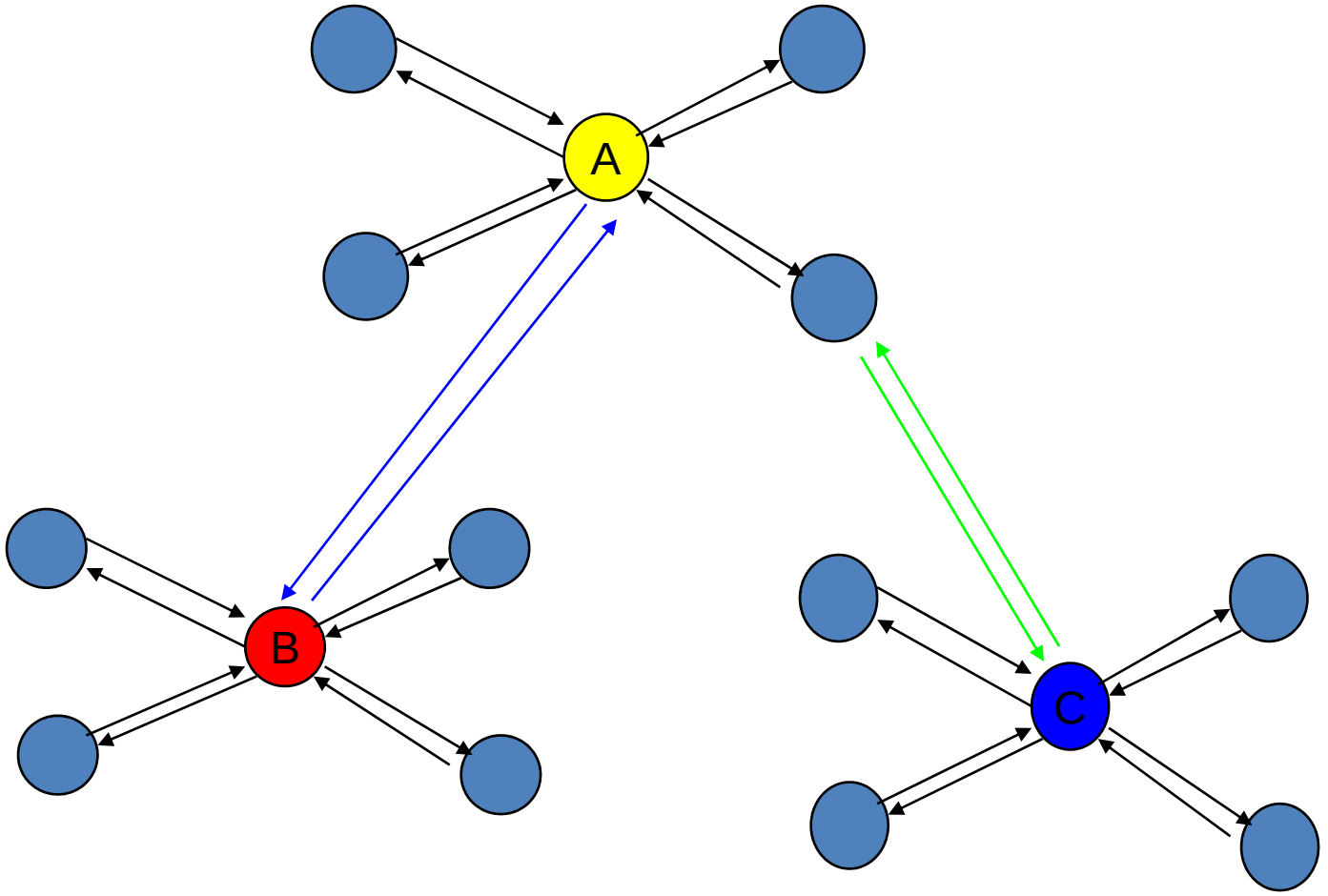
Link exchange web page



Link exchange



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)

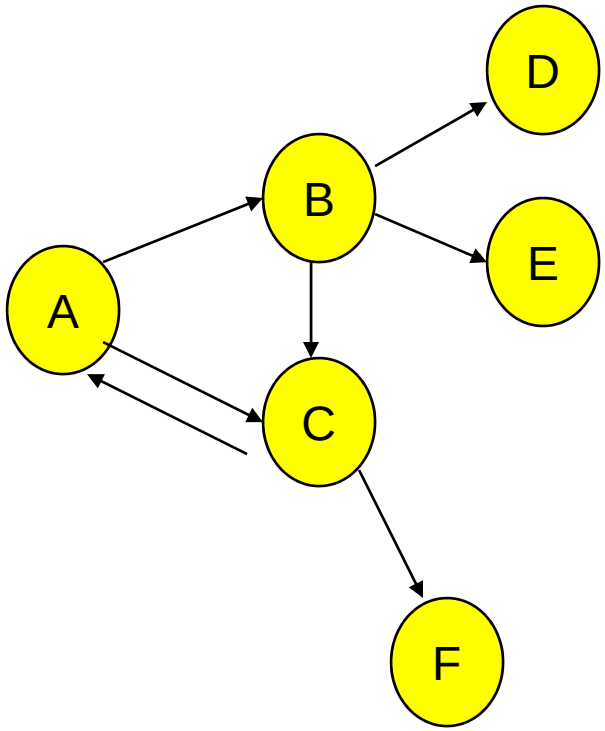
Transition Probability Matrix

- Transition probability matrix at step t

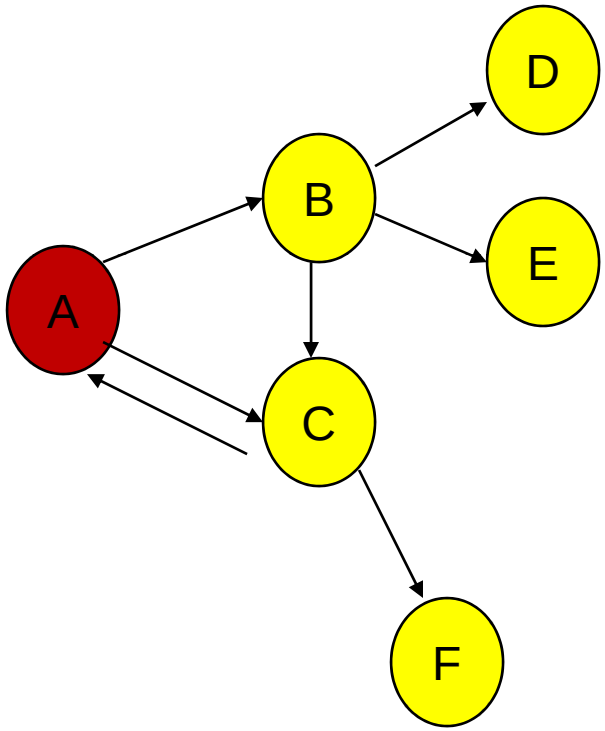
$$p^{t+1} = \frac{1}{2}\left(I + AD^{-1}\right)p^{t}$$

- 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

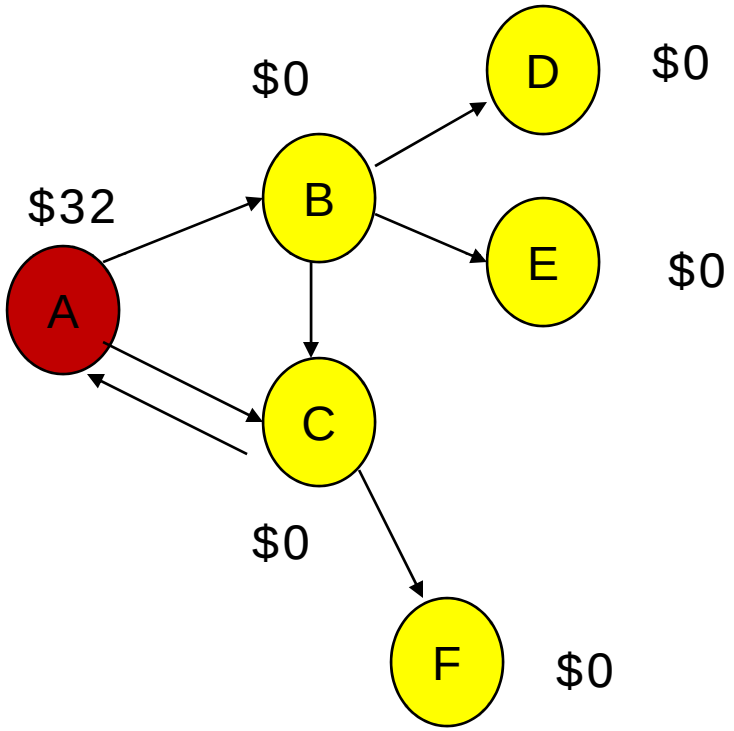


Resource redistribution



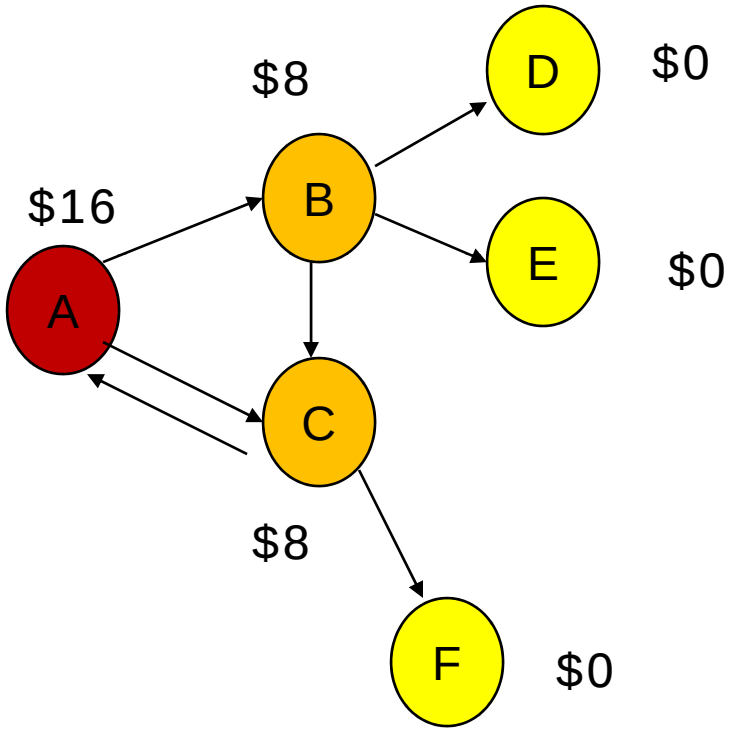
Spam seed

Resource redistribution



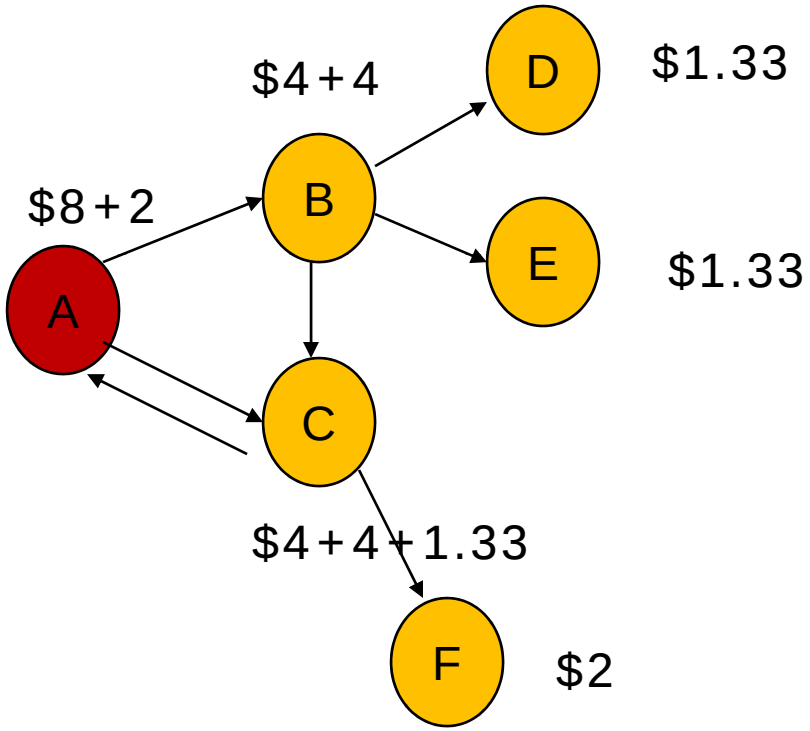
Initialization

Resource redistribution



First iteration

Resource redistribution



Second iteration

Biased Random Walk

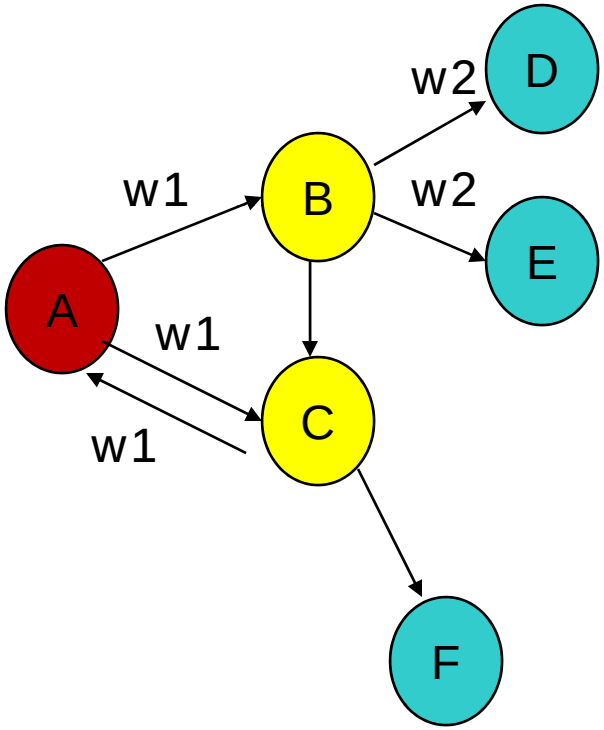
- Bias the random walk to be local
  - Truncation (threshold, percentile)
    - Set very small probabilities to zero
  - Decayed random walk
    - Contain random walk in the neighborhood of the seed set
    - Exponential drop with distance from the seeds
  - White lists and black lists
    - Known good and bad sites
  - Renormalization
    - Compensate for leaked probability

$$p^t[i] = p^t[i] \times \gamma[i]$$

$$\gamma[i] = 2^{-\delta(i)}$$

- The random walk converges after a few iterations

Decayed version



$w1 = w2 = w2$

$w1 = 2w2$

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)

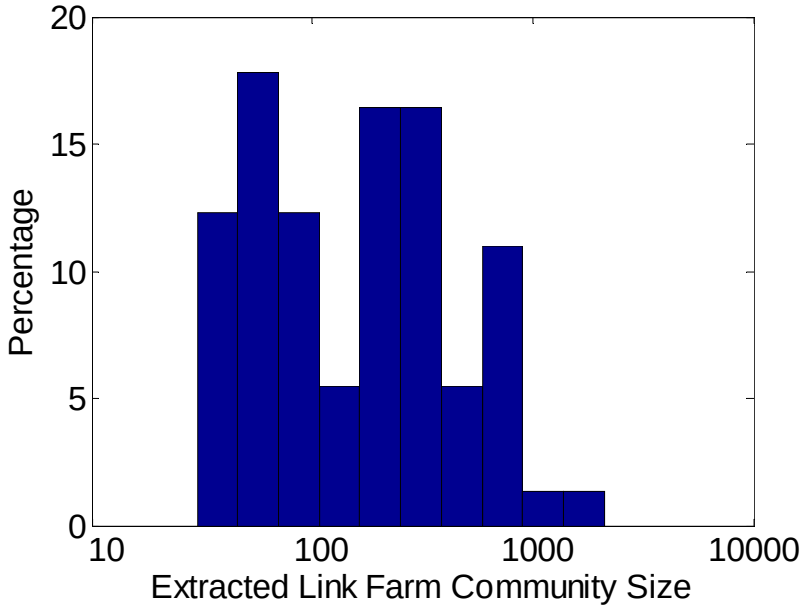
- Directed, Inverted, and Undirected random walks
  - Each seed is processed separately
  - Simulated till convergence (10-30 iterations - a few seconds)

Evaluating Spam Communities

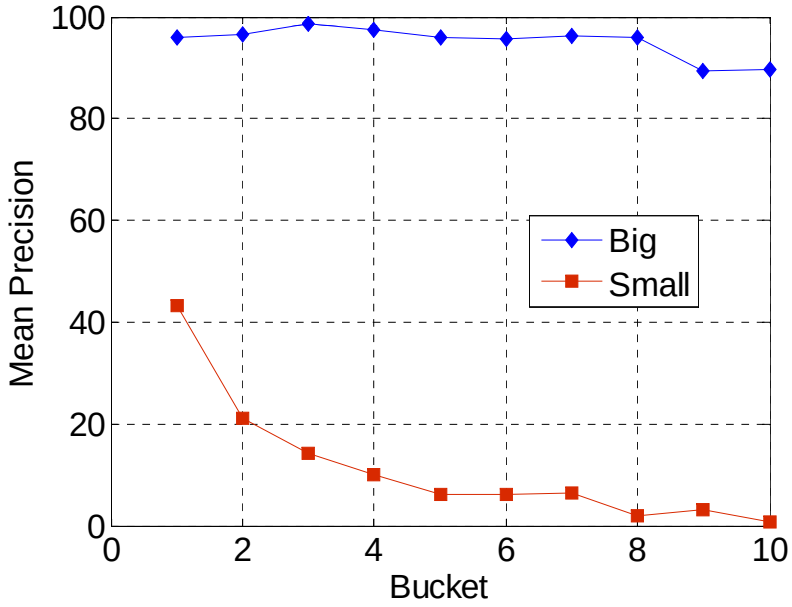
- Sort all the output nodes using on their convergence probability
- Distribute them into ten buckets based on percentiles
  - Bucket 1 has the top 10 percent of the nodes with the highest value
- From each bucket, randomly choose three nodes for manual evaluation.
  - For link farm, evaluate whether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

Link Farm Results

Spam community sizes (mean = 268)

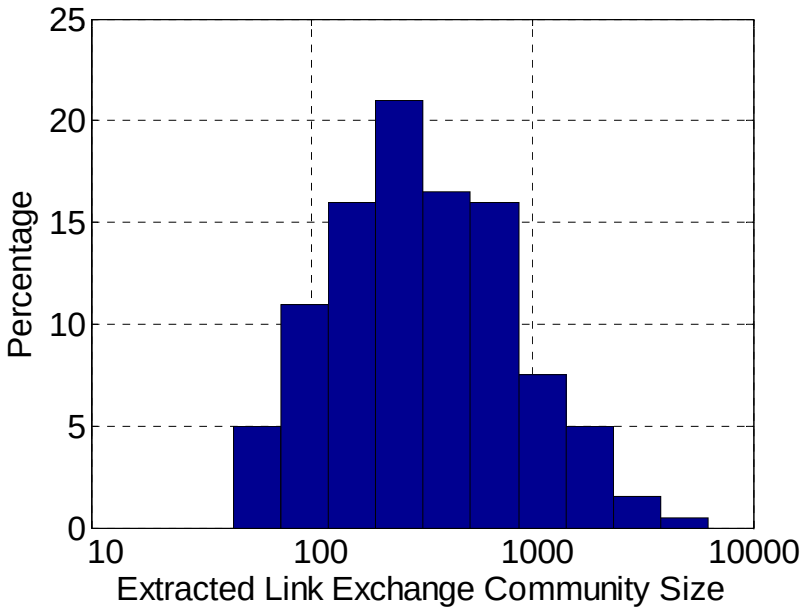


Precision



Link Exchange Results

Spam community sizes (mean = 513)



Precision