TRAWLING THE WEB FOR EMERGING CYBER-COMMUNITIES

Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins
IBM Almaden Research Center

Presented by Xiaoguang Qi
Overview

• Web communities: groups of individuals who share a common interest, together with the web pages most popular among them
  – Explicitly-defined communities: mature, easy to find, 10,000
  – Implicitly-defined communities: emerging, focused on a fine level, 100,000

• Goal: finding implicit (emerging) web communities

• Why?
  – They provide valuable and reliable information resources
  – They represent the sociology of the web
  – Studying them helps to target advertising at a precise level
Basic Idea and Challenge

• Intuition: web communities are characterized by dense directed bipartite graphs
• Each web community contains at least one core, where a core is a \((i, j)\) complete bipartite
• Basic idea: finding emerging web communities by enumerating complete bipartite in the web graph
• Challenge: scalability
  – Date source: 200 million web pages
  – Co-citation relation is very large
  – Even link-to relation is too large to fit in memory
• Goals
  – Stream input
  – Execution time: linear
First steps

• Fans and centers

• Finding potential fans:
  – A potential fan page has links to at least 6 different websites

• Detecting duplications
  – Mirrored fans and centers are inevitable
  – Could be a potential problem here
  – Detect duplications with shingling method, Broder et. al.

• Pruning centers by in-degree
  – Delete all pages that have in-degree larger than $k$
  – $k = 50$
Iterative Pruning

• For (i,j) cores, iteratively
  – Prune potential fans with out-degree smaller than j
  – Delete associated edges
  – Prune potential centers with in-degree smaller than i
  – Delete associated edges

• Implementation
  – Two sorted lists of the edges, one by source, the other by destination
  – Retain in memory a sorted list of edges pruned in each iteration
  – No need to sort the lists after each iteration
Inclusion-Exclusion Pruning

- At every step, we either eliminate a page from contention, or discover an \((i,j)\) core.
- Let \(\{c_1, c_2, \ldots, c_j\}\) be the centers pointed to by \(x\), \(N(c_t)\) denote the set of fans point to \(c_t\), then

\[
x is part of a core \iff \bigcap_{t=1}^{j} N(c_t) \geq i
\]
Inclusion-Exclusion Pruning (Cont.)

- **Implementation**
  - Still two sorted lists of the edges, one by source (L1), the other by destination (L2)
  - Each fan x has a set $S(x)$, which initialized to the complete set
  - Scan L1, find fan x with out-degree exactly j
  - In memory, index edges associated with x by destination (R)
  - Repeat for as many of the fans as the memory can hold the index
  - Stream through L2, for each destination y, check if it is in R
  - If yes, for each $(x', y)$ in R, $S(x') = S(x') \cap N(y)$
  - Check if $S(x')$ has size at least j
Core Generation and Filtering

- Cores are output during inclusion-exclusion pruning
- Next, nepotistic cores are filtered away
  - A nepotistic core is one where some of the fans in the core come from the same web site
  - Why? Fans from the same web site may be intentionally established by the same entity
- Finally, enumerate all cores in the graph
  - Fix j
  - Start with all (1, j) cores
  - Construct all (2, j) cores by checking every fan which also cites any center in another (1, j) core
  - Continue
Evaluation and Conclusion

• Manual evaluation, 400 communities are randomly selected

• Fossilization:
  – Fossil: all of fan pages of a community do not exist
  – 70% of the 400 communities were still alive

• Reliability
  – Only 4% are coincidental core

• Recoverability

• Quality
  – 29% were not in Yahoo today
  – For those appear in Yahoo, average level is 4.5
Critiques

- **Good points**
  - Defensive style
    - one may argue …, the reason why we do this is that …

- **Things can be improved**
  - The notion of a web site is too coarse
    - “allentowngasprices.com”, “pittsburghgasprices.com”
  - Evaluation seems not sound enough
    - “29% of the sampled communities were not in Yahoo”
      - Maybe they no longer exist
  - Some typos
  - Some confusing expressions (at least to me)