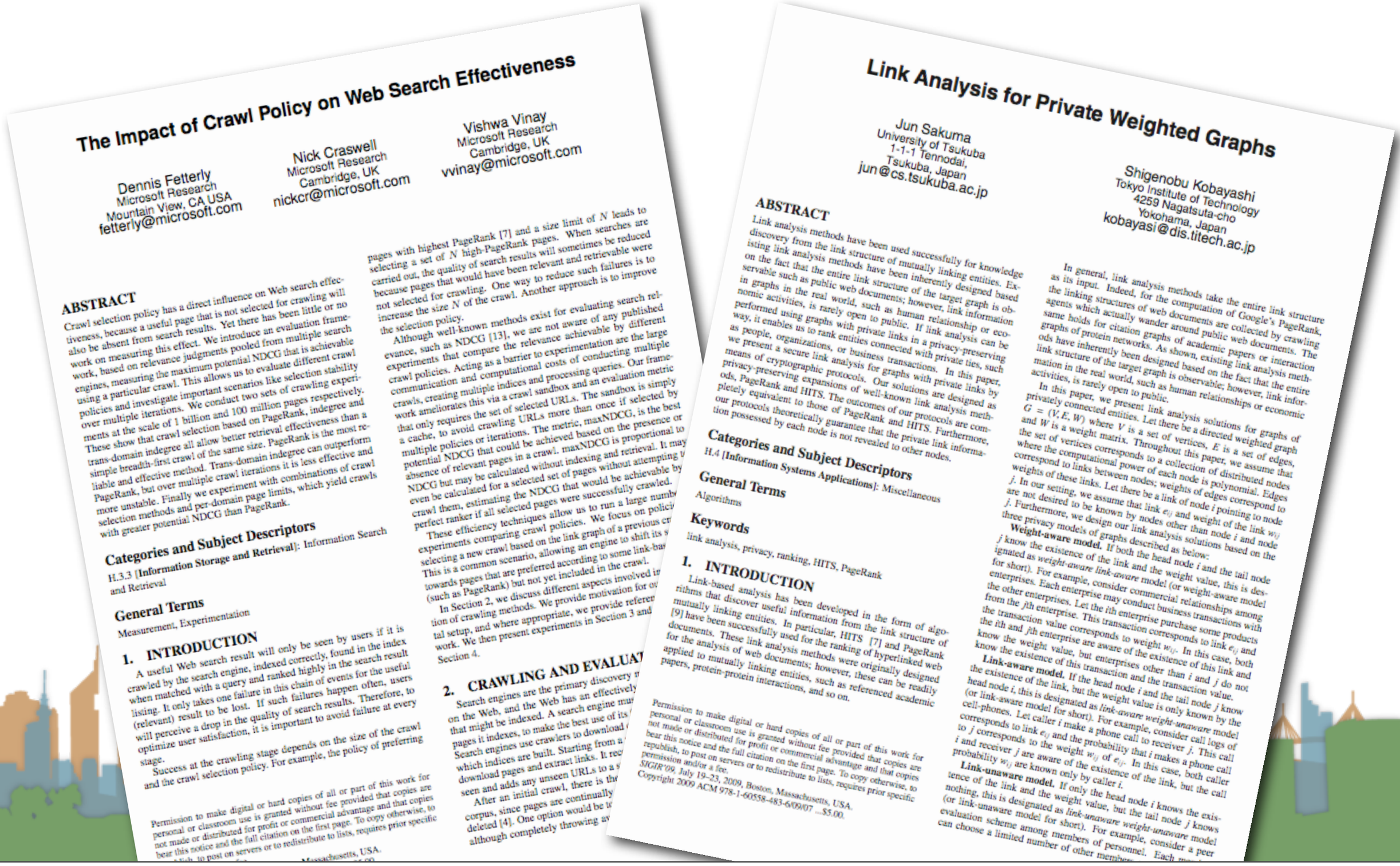
Advances in Link Analysis

Introduction to Information Retrieval
CS 221
Donald J. Patterson

Advances in Link Analysis

"The Impact of Crawl Policy on Web Search Effectiveness" by Fetterly, Craswell, Vinay SIGIR2009

"Link Analysis for Private Weighted Graphs" by Sakuma, Kobayashi SIGIR2009



The Impact of Crawl Policy on Web Search Effectiveness

Dennis Fetterly
Microsoft Research
Mountain View, CA USA
fetterly@microsoft.com

Nick Craswell
Microsoft Research
Cambridge, UK
nickcr@microsoft.com

Vishwa Vinay
Microsoft Research
Cambridge, UK
vvinay@microsoft.com

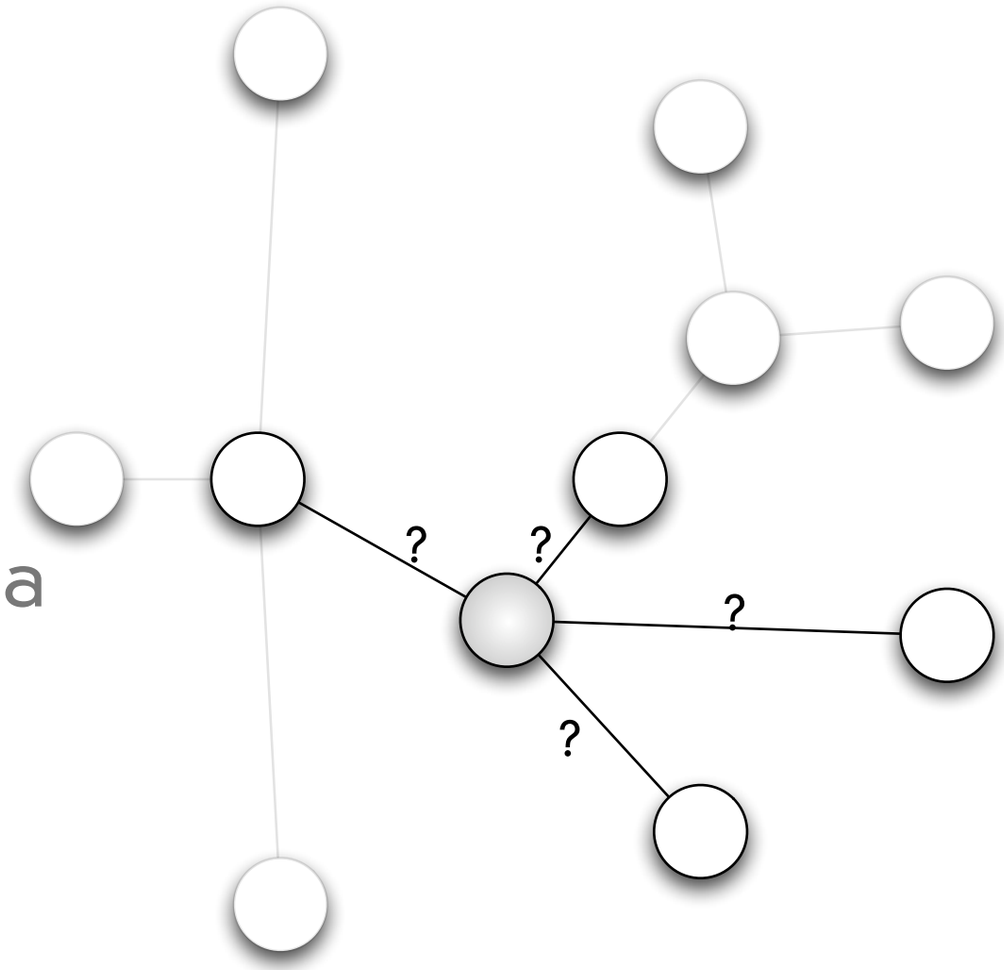
Link Analysis for Private Weighted Graphs

Jun Sakuma
University of Tsukuba
1-1-1 Tennodai,
Tsukuba, Japan
jun@cs.tsukuba.ac.jp

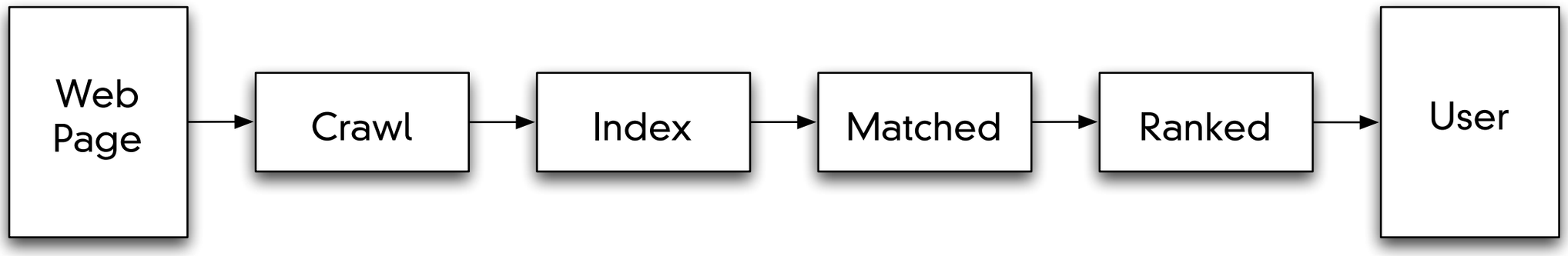
Shigenobu Kobayashi
Tokyo Institute of Technology
4259 Nagatsuta-cho
Yokohama, Japan
kobayasi@dis.titech.ac.jp

The Impact of Crawl Policy on Web Search Effectiveness

- Crawl Selection Policy
  - A good page, not crawled, is not returned for a query
  - Deciding what to crawl can make a difference with finite resources

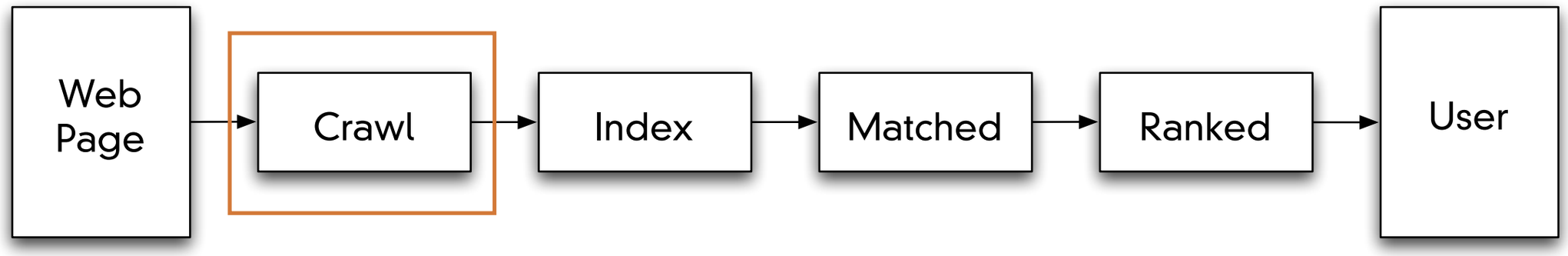


The Impact of Crawl Policy on Web Search Effectiveness



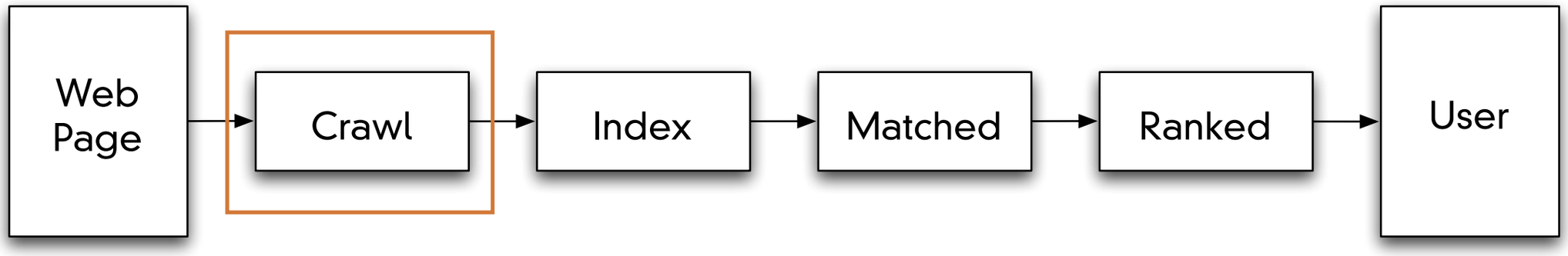
- For a web page to get to a user in a search result many things must go well
- A failure results in worse quality and unhappy users
- Instead of holding the corpus constant and testing retrieval algorithms, this paper varies the corpus through crawl strategies

The Impact of Crawl Policy on Web Search Effectiveness



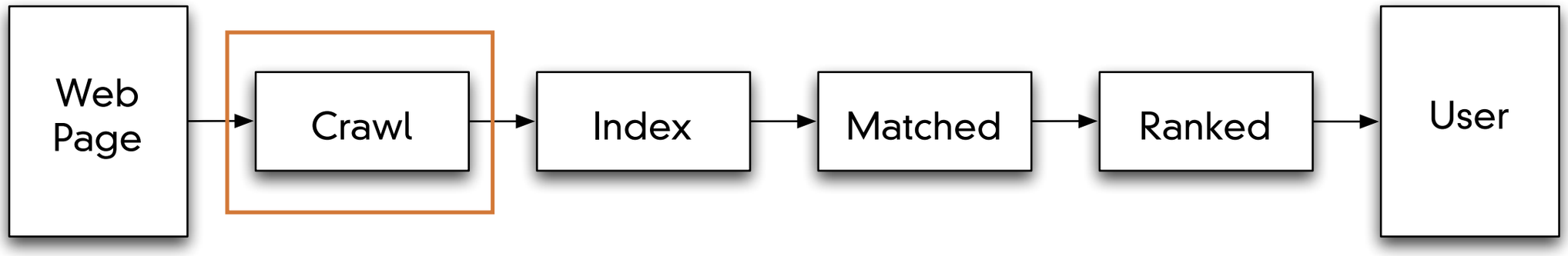
- Successful crawling depends on
  - size of crawl
  - crawl selection policy

The Impact of Crawl Policy on Web Search Effectiveness



- Improving crawling depends on
  - Increasing the size of the crawl, or
  - using a better crawl selection policy

The Impact of Crawl Policy on Web Search Effectiveness



- Improving crawling depends on
  - Increasing the size of the crawl, or
  - using a better crawl selection policy

The Impact of Crawl Policy on Web Search Effectiveness

- This paper
  - Provides a metric to compare crawls
  - evaluates 4 different crawl policies
    - one-time
    - incrementally

The Impact of Crawl Policy on Web Search Effectiveness

- A naive approach to crawling would be to start from a seed set every time
- A smarter approach is to crawl based on quality metrics derived from the last crawl

The Impact of Crawl Policy on Web Search Effectiveness

- 4 crawl selection policies
  - breadth first search (baseline)
  - indegree
  - trans-domain indegree
  - highest PageRank
- Not done:
  - focussing crawl based on queries with lame results

The Impact of Crawl Policy on Web Search Effectiveness

- Options for evaluating a crawl
  - Efficiency
    - politeness
    - queueing
    - data structures
    - resource allocation
  - "Goodness" of collected corpus

The Impact of Crawl Policy on Web Search Effectiveness

- Goodness can be measure by “RankMass”
  - Like precision over PageRank
  - What % of top N pages did you crawl?
  - Requires a big crawl for ground truth
  - Requires crawler
- Alternative is Normalized Discounted Cumulative Gain (NDCG)

The Impact of Crawl Policy on Web Search Effectiveness

- metric: maxNDCG
  - eliminates variation based on retrieval algorithm
  - several thousand test queries
    - 10,570 sampled queries (Microsoft)
  - maxNDCG = "how many of the web pages returned by 3 large web search engines are in your crawl set?" weighted by human judged relevance

The Impact of Crawl Policy on Web Search Effectiveness

- metric: NDCG@K
  - G(j) is user rating of document at position j

$$\sum_{j=1}^{K} \left(\frac{G(j)}{log(1+j)} \right)$$

The Impact of Crawl Policy on Web Search Effectiveness

- metric: click utility
  - sum of clicks on pages in crawled corpus
  - requires search engine support (Microsoft)
  - not weighted by rank in query results
    - (multiple queries)
  - like RankMass for clicks

The Impact of Crawl Policy on Web Search Effectiveness

- incremental metric: Jaccard coefficent
  - High Jaccard coefficient means not many pages changed between crawls

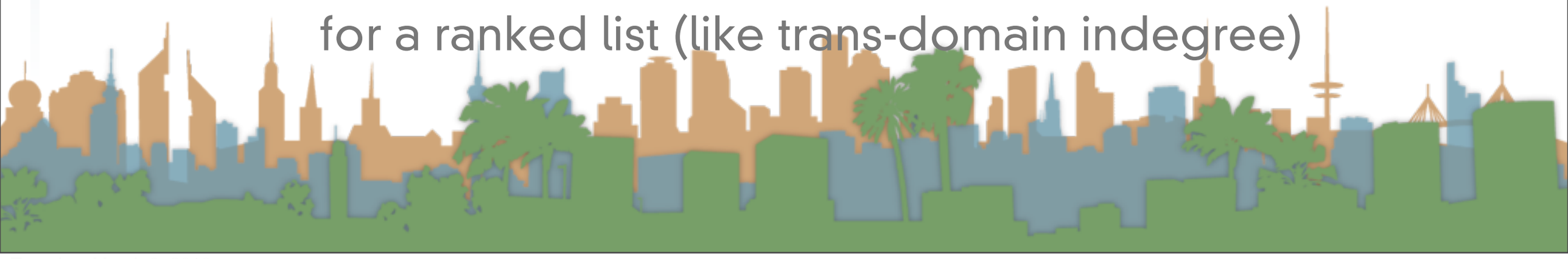
$$\frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}$$

The Impact of Crawl Policy on Web Search Effectiveness

- incremental metric: churn
  - How many pages came and went during the iterative crawls
  - Users care because it makes the search engine look stable

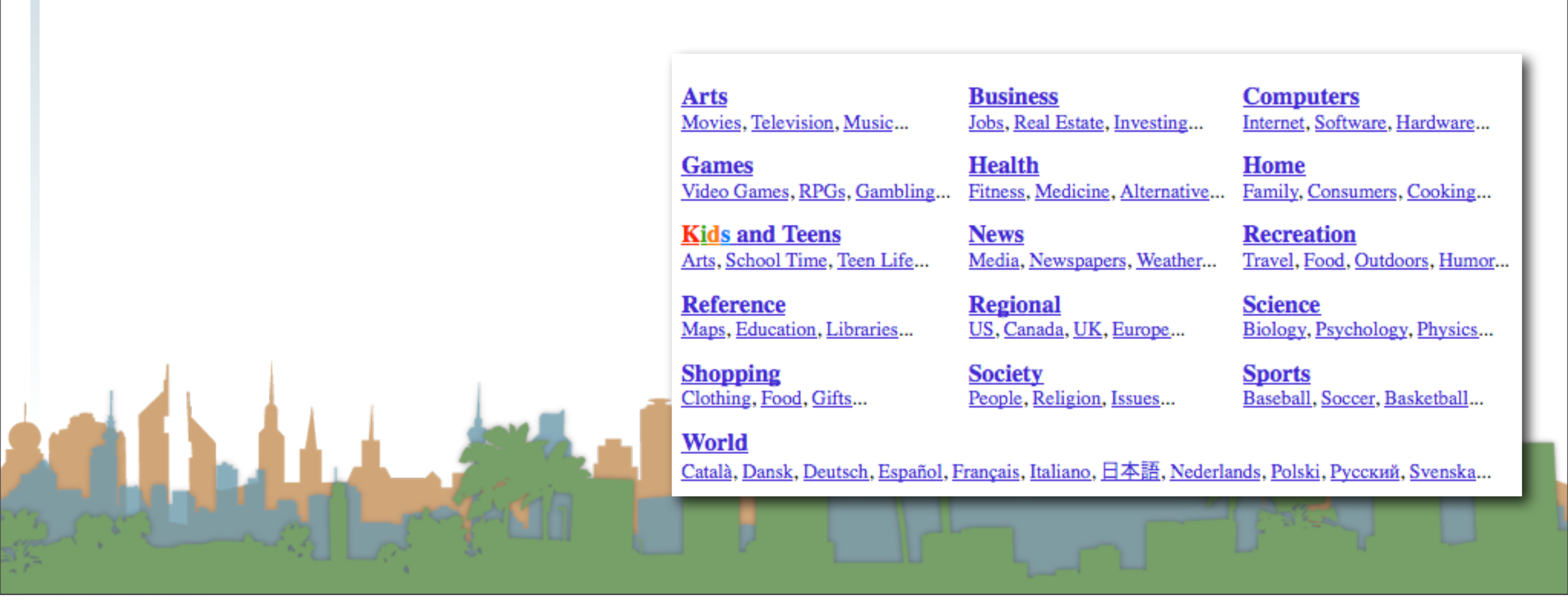
The Impact of Crawl Policy on Web Search Effectiveness

- Systems issues
  - Hard to compare crawls when the Internet is dynamic
  - Real-time events are irreproducible
    - network transience
  - For this experiment
    - Sandbox crawl (basically a big cache)
      - First time a URL is requested it is cached.
      - Time is not being evaluated
    - When required equal scores are randomly selected for a ranked list (like trans-domain indegree)



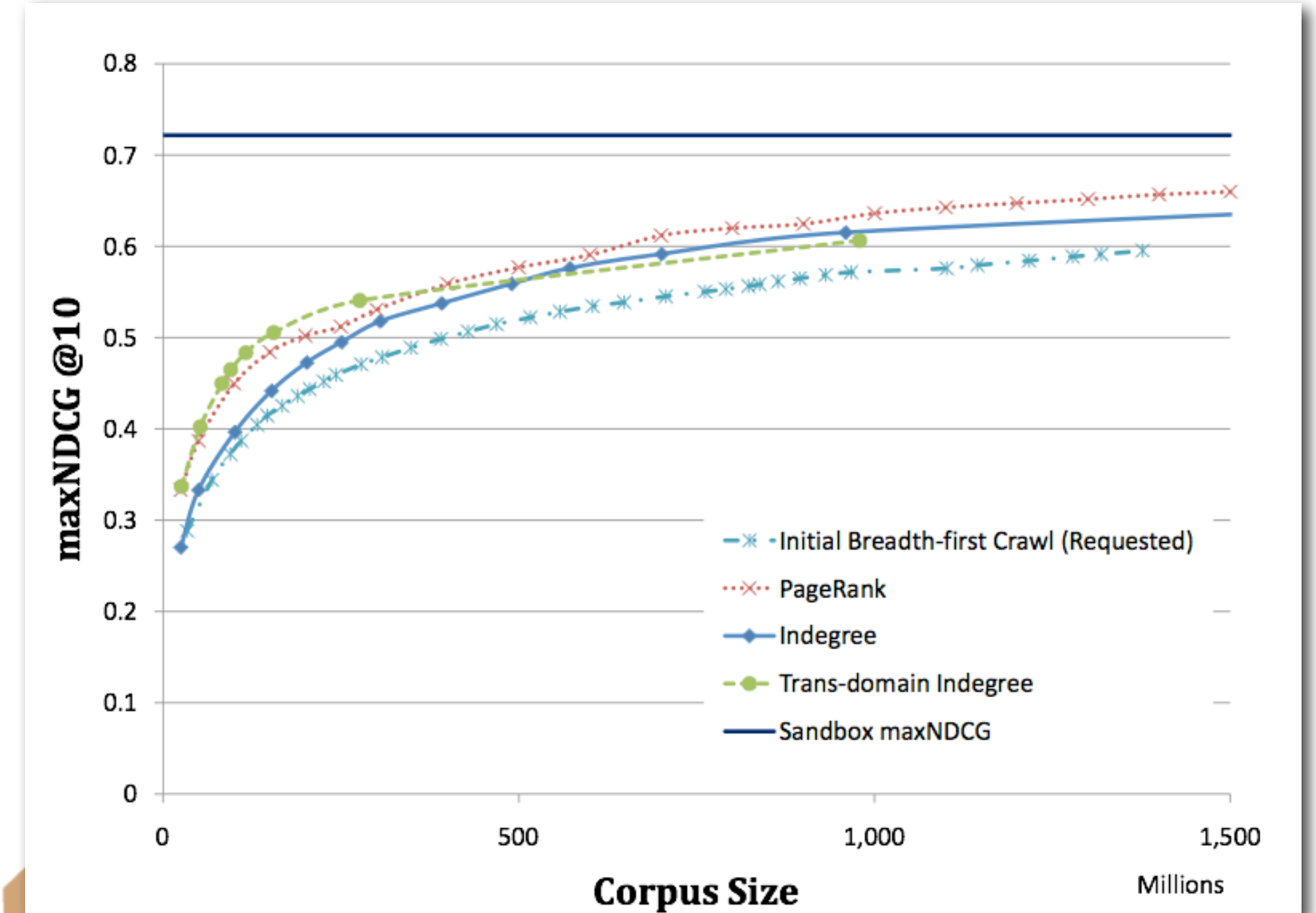
The Impact of Crawl Policy on Web Search Effectiveness

- Experiments
  - Baseline
    - breadth-first crawl starting at open-directory
    - 930,320,010 filtered pages recovered
    -



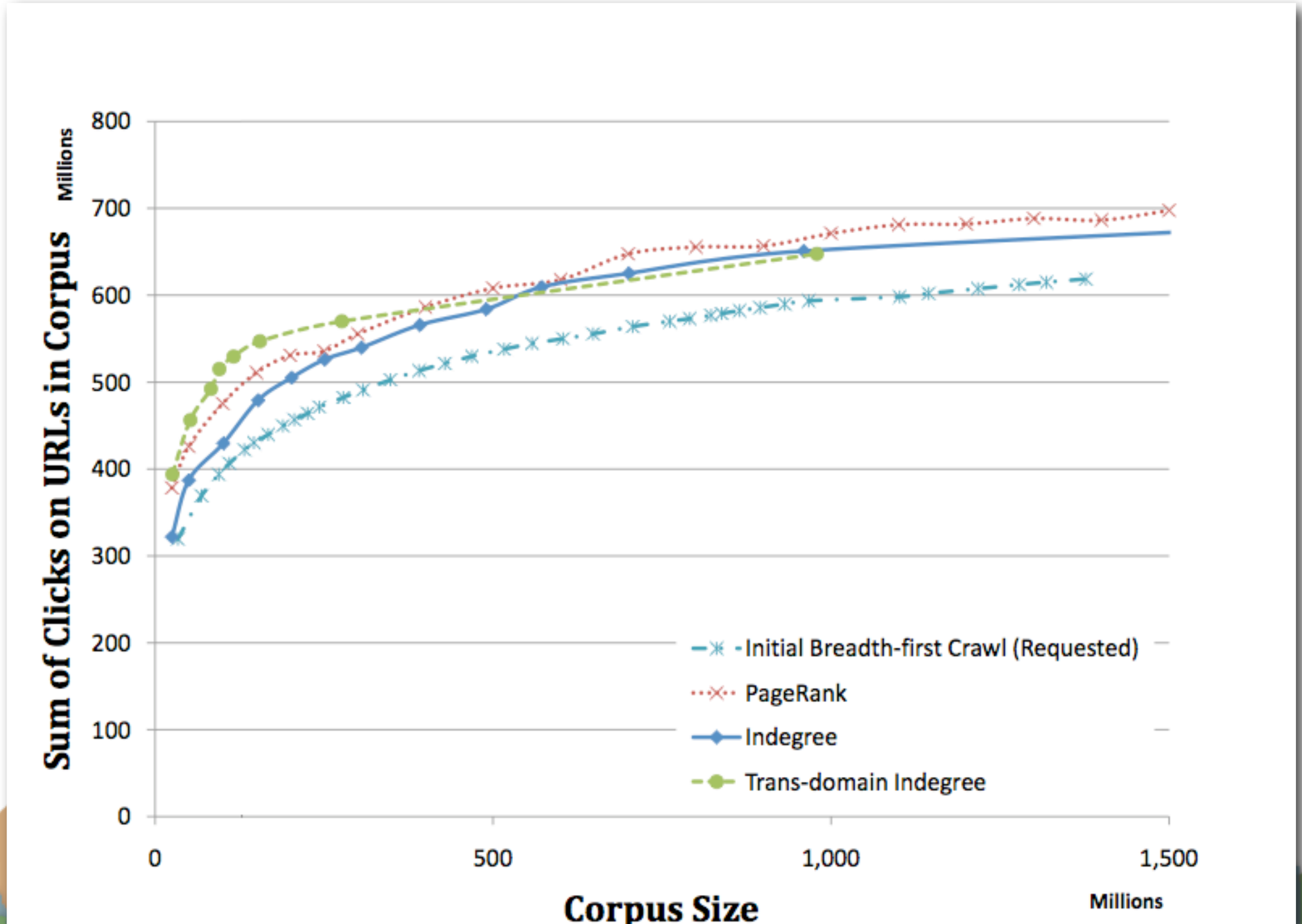
The Impact of Crawl Policy on Web Search Effectiveness

- Experiments - Single Crawl



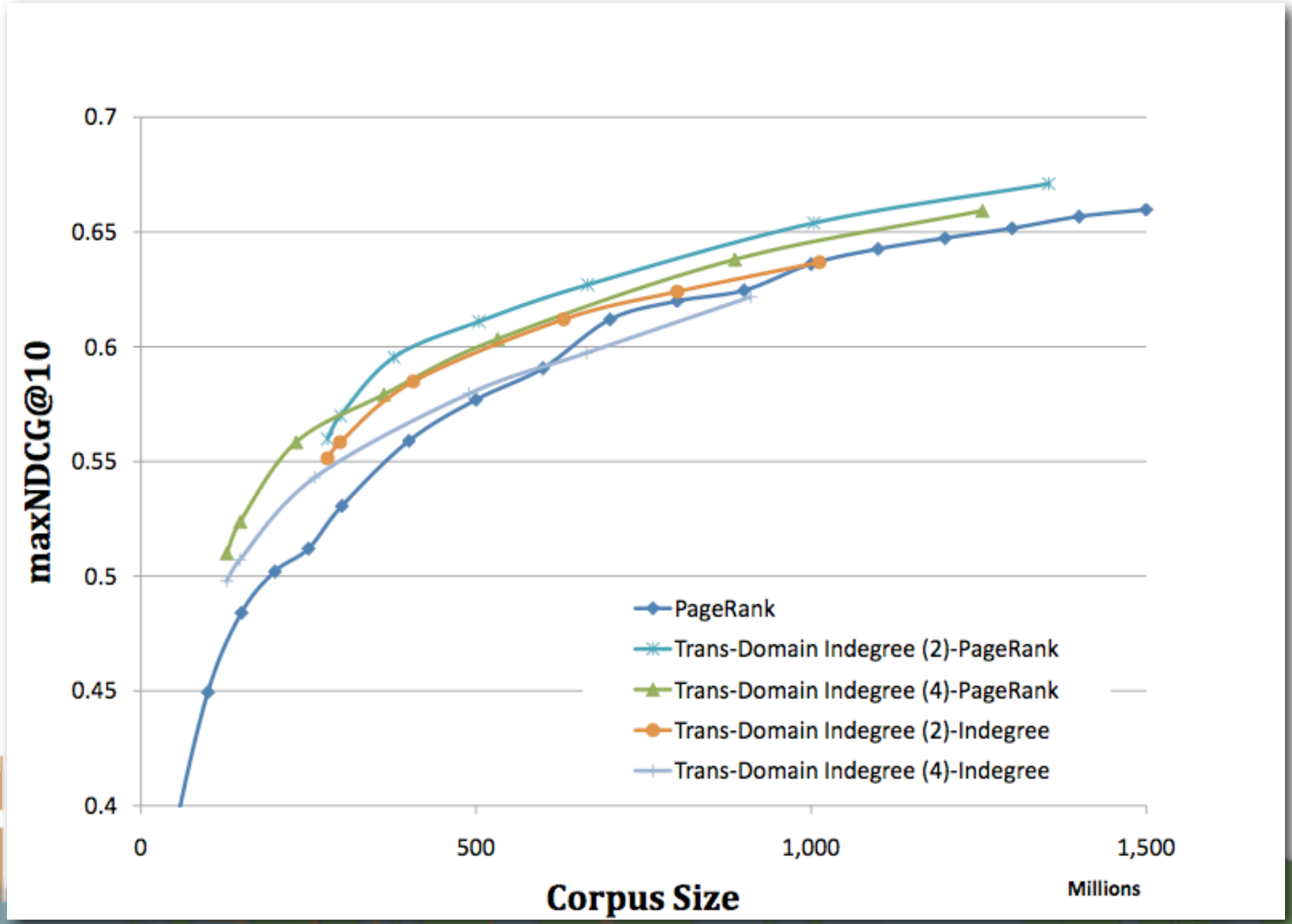
The Impact of Crawl Policy on Web Search Effectiveness

- Experiments - Single Crawl



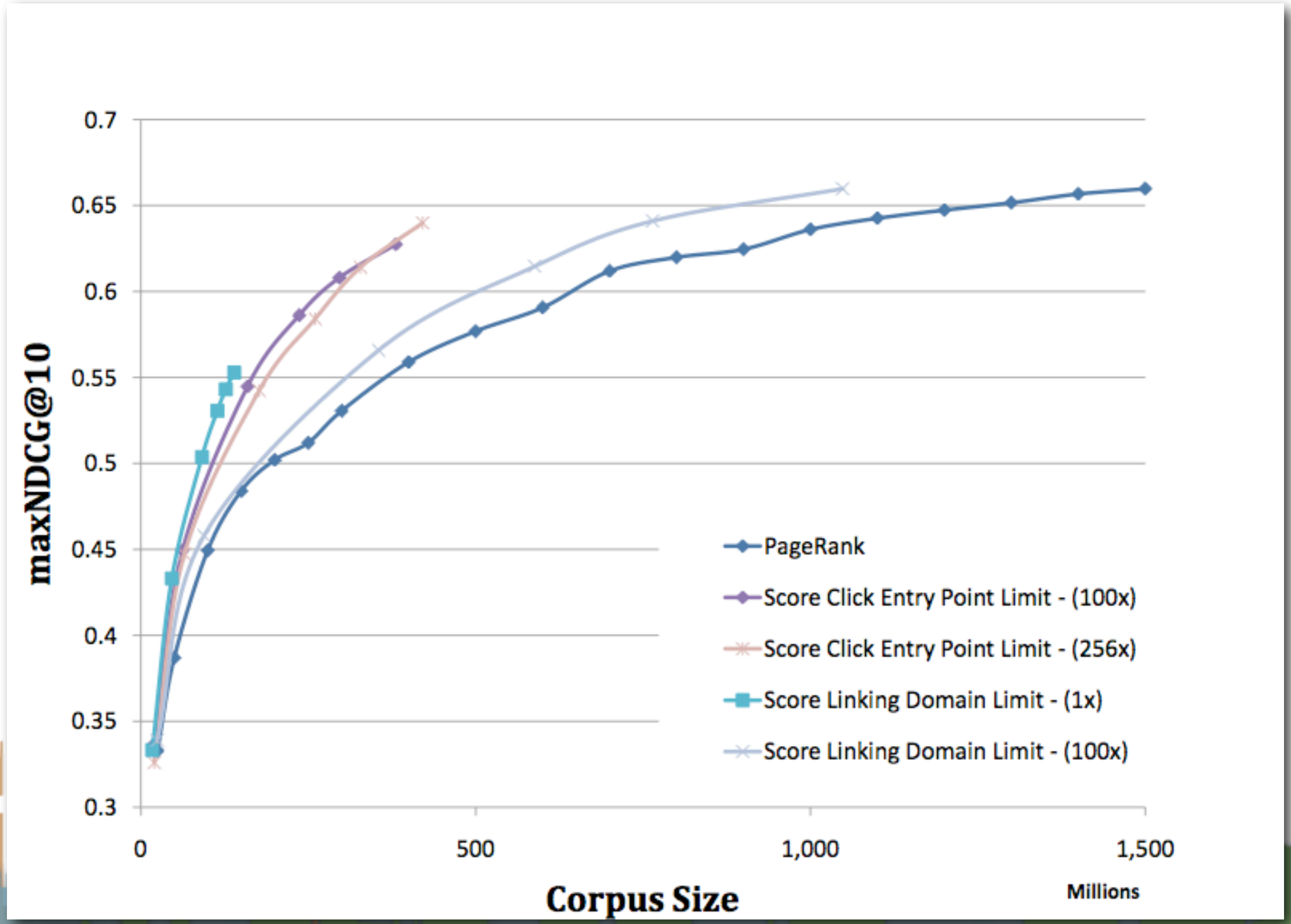
The Impact of Crawl Policy on Web Search Effectiveness

- Experiments - Hybrid Crawl



The Impact of Crawl Policy on Web Search Effectiveness

- Experiments - Domain Limits on PageRank Crawl



The Impact of Crawl Policy on Web Search Effectiveness

- Experiments - Iterative Behavior

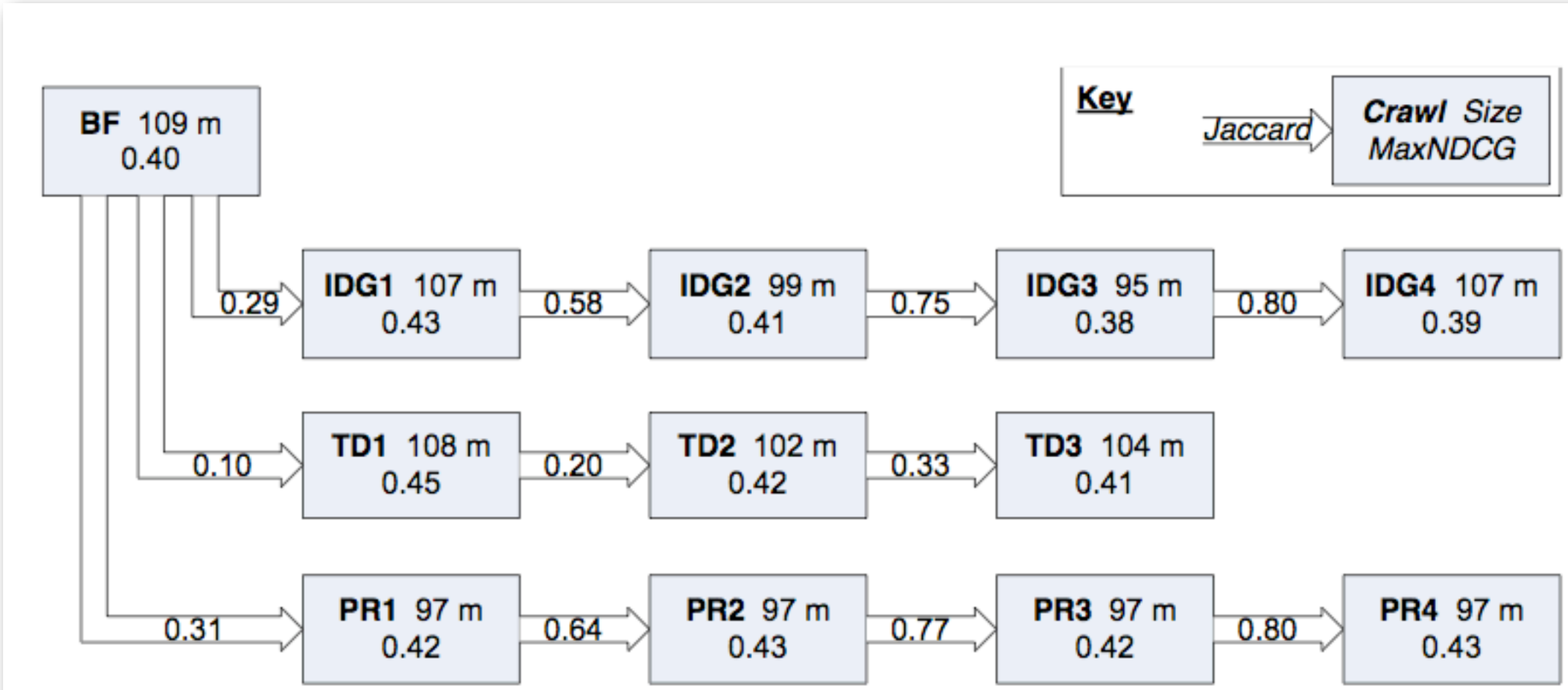


Figure 4: Iterative selection experiments, showing the size and maxNDCG of requested selections, and the Jaccard similarity between iterations.

The Impact of Crawl Policy on Web Search Effectiveness

- Experiments - Iterative Set Similarity

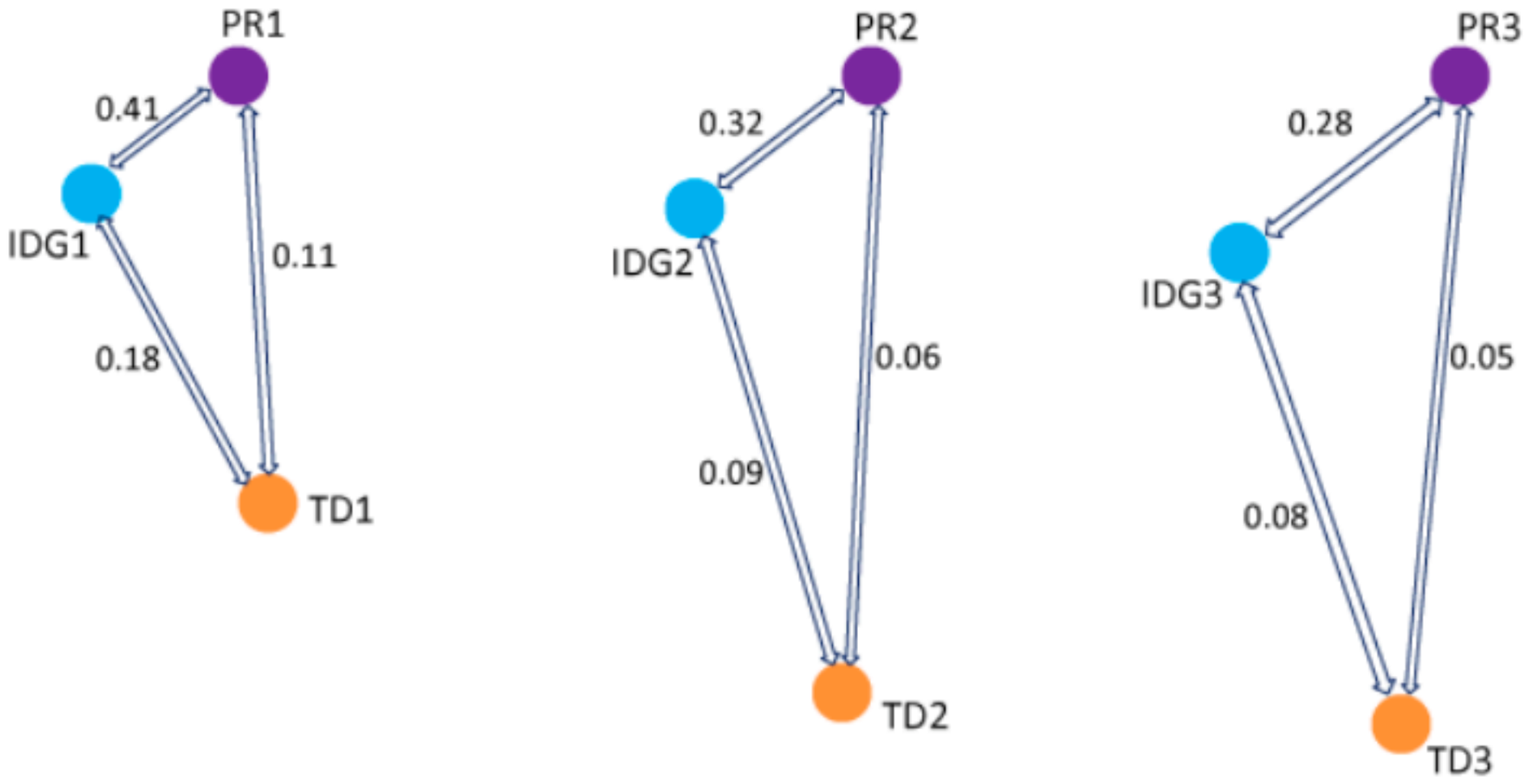
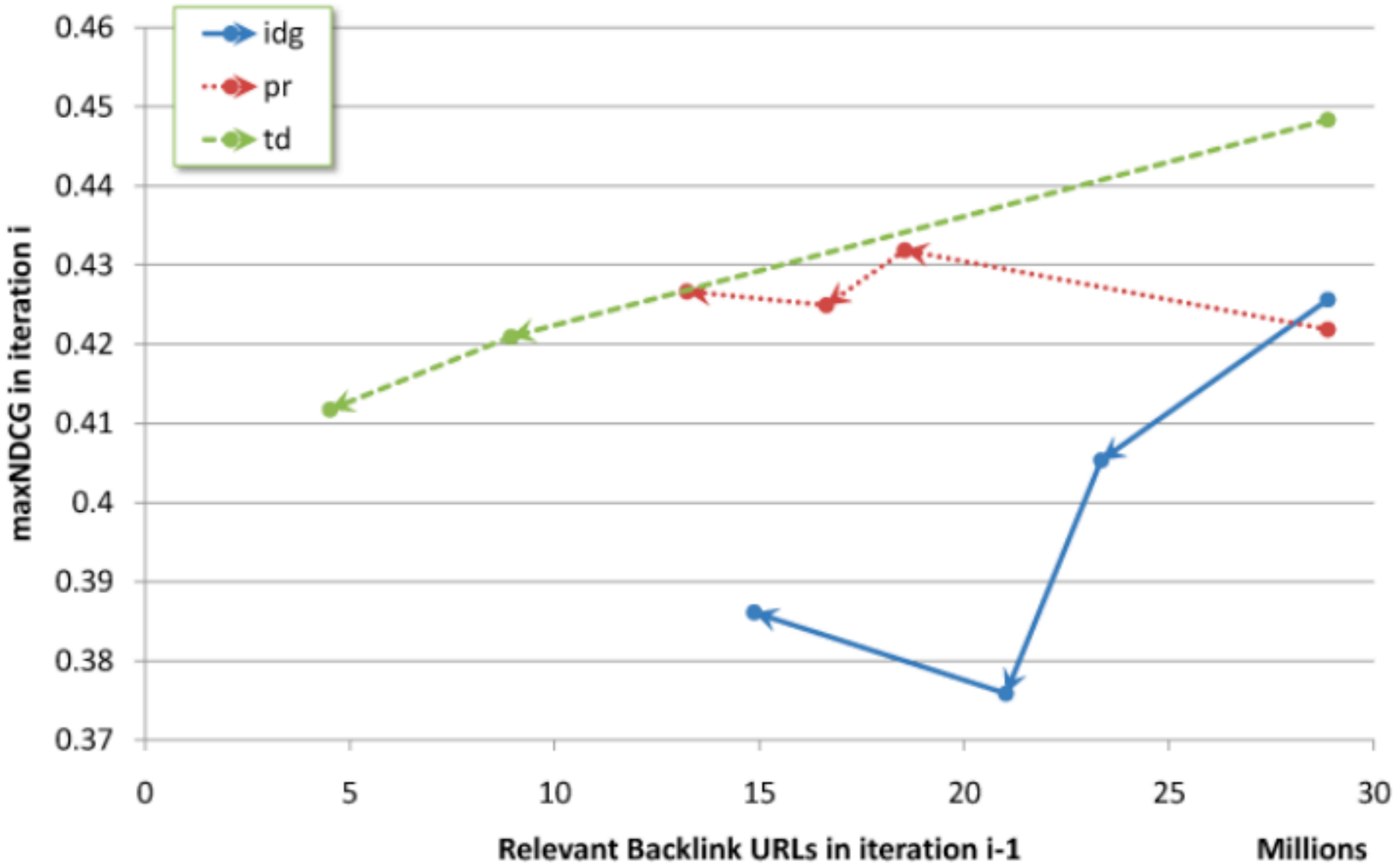


Figure 5: Jaccard similarity between corpora selected by different policies at different iterations.

The Impact of Crawl Policy on Web Search Effectiveness

- Experiments - Iterative Self-Similarity



The Impact of Crawl Policy on Web Search Effectiveness

- Experiments - Iterative Stability

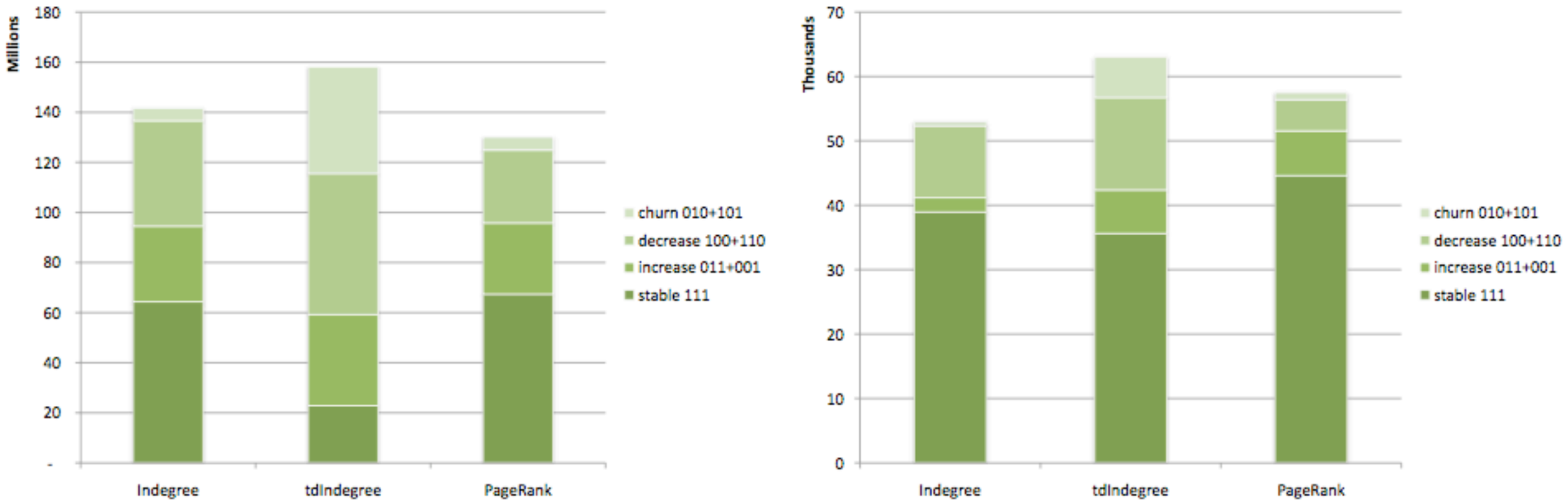


Figure 6: Stability of crawl policies, over three iterations. Left: All URLs. Right: Relevant URLs

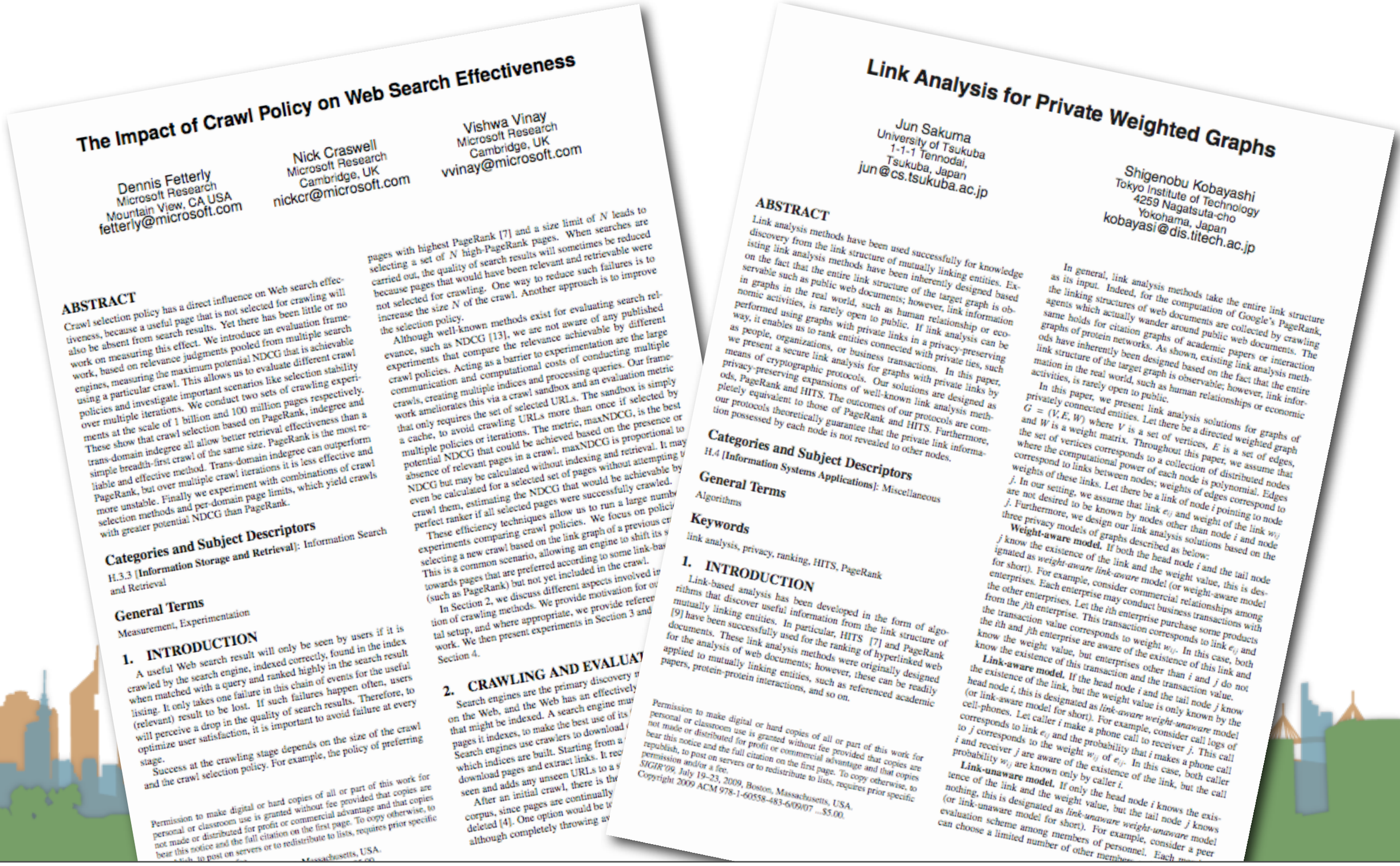
The Impact of Crawl Policy on Web Search Effectiveness

- Take Away
  - Use a combination of TD and PageRank to guide crawl
  - Don’t overcrawl one domain
  - maxNDCG is a measure comparable to user click utility

Advances in Link Analysis

"The Impact of Crawl Policy on Web Search Effectiveness" by Fetterly, Craswell, Vinay SIGIR2009

"Link Analysis for Private Weighted Graphs" by Sakuma, Kobayashi SIGIR2009



The Impact of Crawl Policy on Web Search Effectiveness

Dennis Fetterly
Microsoft Research
Mountain View, CA USA
fetterly@microsoft.com

Nick Craswell
Microsoft Research
Cambridge, UK
nickcr@microsoft.com

Vishwa Vinay
Microsoft Research
Cambridge, UK
vvinay@microsoft.com

Link Analysis for Private Weighted Graphs

Jun Sakuma
University of Tsukuba
1-1-1 Tennodai,
Tsukuba, Japan
jun@cs.tsukuba.ac.jp

Shigenobu Kobayashi
Tokyo Institute of Technology
4259 Nagatsuta-cho
Yokohama, Japan
kobayasi@dis.titech.ac.jp