

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

ABSTRACT

Crawl selection policy has a direct influence on Web search effectiveness, because a useful page that is not selected for crawling will also be absent from search results. Yet there has been little or no work on measuring this effect. We introduce an evaluation framework, based on relevance judgments pooled from multiple search engines, measuring the maximum potential NDCG that is achievable using a particular crawl. This allows us to evaluate different crawl policies and investigate important scenarios like selection stability over multiple iterations. We conduct two sets of crawling experiments at the scale of 1 billion and 100 million pages respectively. These show that crawl selection based on PageRank, indegree and trans-domain indegree all allow better retrieval effectiveness than a simple breadth-first crawl of the same size. PageRank is the most reliable and effective method. Trans-domain indegree can outperform PageRank, but over multiple crawl iterations it is less effective and more unstable. Finally we experiment with combinations of crawl selection methods and per-domain page limits, which yield crawls with greater potential NDCG than PageRank.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms

Measurement, Experimentation

1. INTRODUCTION

A useful Web search result will only be seen by users if it is crawled by the search engine, indexed correctly, found in the index when matched with a query and ranked highly in the search result listing. It only takes one failure in this chain of events for the useful (relevant) result to be lost. If such failures happen often, users will perceive a drop in the quality of search results. Therefore, to optimize user satisfaction, it is important to avoid failure at every stage.

Success at the crawling stage depends on the size of the crawl and the crawl selection policy. For example, the policy of preferring

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 2009 ACM 978-1-60558-483-6/09/07 ...\$5.00.

pages with highest PageRank [7] and a size limit of N leads to selecting a set of N high-PageRank pages. When searches are carried out, the quality of search results will sometimes be reduced because pages that would have been relevant and retrievable were not selected for crawling. One way to reduce such failures is to increase the size N of the crawl. Another approach is to improve the selection policy.

Although well-known methods exist for evaluating search relevance, such as NDCG [13], we are not aware of any published experiments that compare the relevance achievable by different crawl policies. Acting as a barrier to experimentation are the large communication and computational costs of conducting multiple crawls, creating multiple indices and processing queries. Our framework ameliorates this via a crawl sandbox and an evaluation metric that only requires the set of selected URLs. The sandbox is simply a cache, to avoid crawling URLs more than once if selected by multiple policies or iterations. The metric, maxNDCG, is the best potential NDCG that could be achieved based on the presence or absence of relevant pages in a crawl. maxNDCG is proportional to NDCG but may be calculated without indexing and retrieval. It may even be calculated for a selected set of pages without attempting to crawl them, estimating the NDCG that would be achievable by a perfect ranker if all selected pages were successfully crawled.

These efficiency techniques allow us to run a large number of experiments comparing crawl policies. We focus on policy selection based on the link graph of a previous crawl. This is a common scenario, allowing an engine to shift its focus towards pages that are preferred according to some link-based metric (such as PageRank) but not yet included in the crawl.

In Section 2, we discuss different aspects involved in the selection of crawling methods. We provide motivation for our experimental setup, and where appropriate, we provide references to other work. We then present experiments in Section 3 and Section 4.

2. CRAWLING AND EVALUATION

Search engines are the primary discovery mechanism on the Web, and the Web has an effectively infinite number of pages that might be indexed. A search engine must select a subset of pages to index, to make the best use of its resources. Search engines use crawlers to download pages and extract links. It is important to select which indices are built. Starting from a set of seed URLs, which indices are built. Starting from a set of seed URLs, which indices are built. Starting from a set of seed URLs, which indices are built.

After an initial crawl, there is a growing corpus, since pages are continually added and deleted [4]. One option would be to delete completely throwing away

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

ABSTRACT

Link analysis methods have been used successfully for knowledge discovery from the link structure of mutually linking entities. Existing link analysis methods have been inherently designed based on the fact that the entire link structure of the target graph is observable such as public web documents; however, link information in graphs in the real world, such as human relationship or economic activities, is rarely open to public. If link analysis can be performed using graphs with private links in a privacy-preserving way, it enables us to rank entities connected with private ties, such as people, organizations, or business transactions. In this paper, we present a secure link analysis for graphs with private links by means of cryptographic protocols. Our solutions are designed as privacy-preserving expansions of well-known link analysis methods, PageRank and HITS. The outcomes of our protocols are completely equivalent to those of PageRank and HITS. Furthermore, our protocols theoretically guarantee that the private link information possessed by each node is not revealed to other nodes.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

General Terms

Algorithms

Keywords

link analysis, privacy, ranking, HITS, PageRank

1. INTRODUCTION

Link-based analysis has been developed in the form of algorithms that discover useful information from the link structure of mutually linking entities. In particular, HITS [7] and PageRank [9] have been successfully used for the ranking of hyperlinked web documents. These link analysis methods were originally designed for the analysis of web documents; however, these can be readily applied to mutually linking entities, such as referenced academic papers, protein-protein interactions, and so on.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
Copyright 2009 ACM 978-1-60558-483-6/09/07 ...\$5.00.

In general, link analysis methods take the entire link structure as its input. Indeed, for the computation of Google's PageRank, the linking structures of web documents are collected by crawling agents which actually wander around public web documents. The same holds for citation graphs of academic papers or interaction graphs of protein networks. As shown, existing link analysis methods have inherently been designed based on the fact that the entire link structure of the target graph is observable; however, link information in the real world, such as human relationships or economic activities, is rarely open to public.

In this paper, we present link analysis solutions for graphs of privately connected entities. Let there be a directed weighted graph $G = (V, E, W)$ where V is a set of vertices, E is a set of edges, and W is a weight matrix. Throughout this paper, we assume that the set of vertices corresponds to a collection of distributed nodes where the computational power of each node is polynomial. Edges correspond to links between nodes; weights of edges correspond to weights of these links. Let there be a link of node i pointing to node j . In our setting, we assume that link e_{ij} and weight of the link w_{ij} are not desired to be known by nodes other than node i and node j . Furthermore, we design our link analysis solutions based on the three privacy models of graphs described as below:

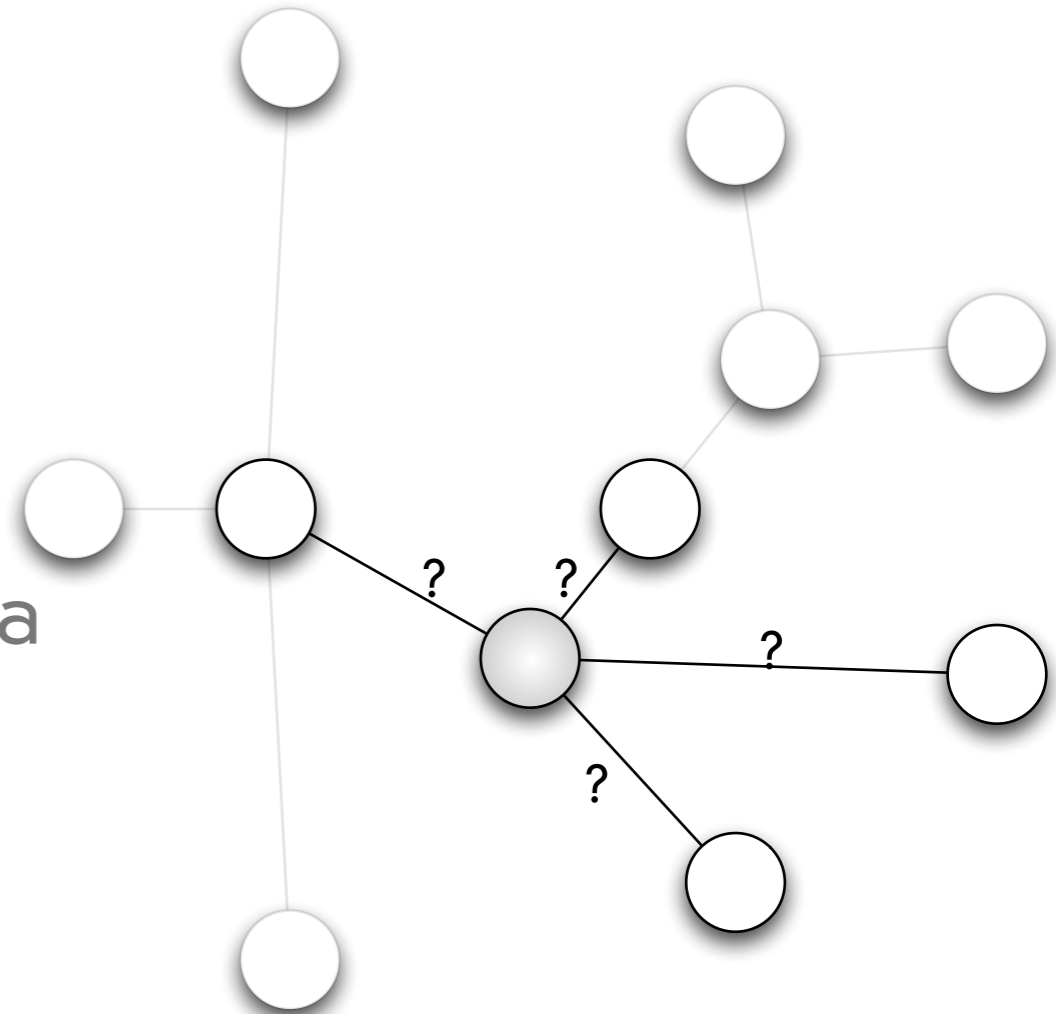
Weight-aware model. If both the head node i and the tail node j know the existence of the link and the weight value, this is designated as *weight-aware link-aware model* (or *weight-aware model* for short). For example, consider commercial relationships among enterprises. Each enterprise may conduct business transactions with the other enterprises. Let the i th enterprise purchase some products from the j th enterprise. This transaction corresponds to link e_{ij} and the transaction value corresponds to weight w_{ij} . In this case, both the i th and j th enterprise are aware of the existence of this link and know the weight value, but enterprises other than i and j do not know the existence of this transaction and the transaction value.

Link-aware model. If the head node i and the tail node j know the existence of the link, but the weight value is only known by the head node i , this is designated as *link-aware weight-unaware model* (or *link-aware model* for short). For example, consider call logs of cell-phones. Let caller i make a phone call to receiver j . This call corresponds to link e_{ij} and the probability that i makes a phone call to j corresponds to the weight w_{ij} of e_{ij} . In this case, both caller i and receiver j are aware of the existence of the link, but the caller probability w_{ij} are known only by caller i .

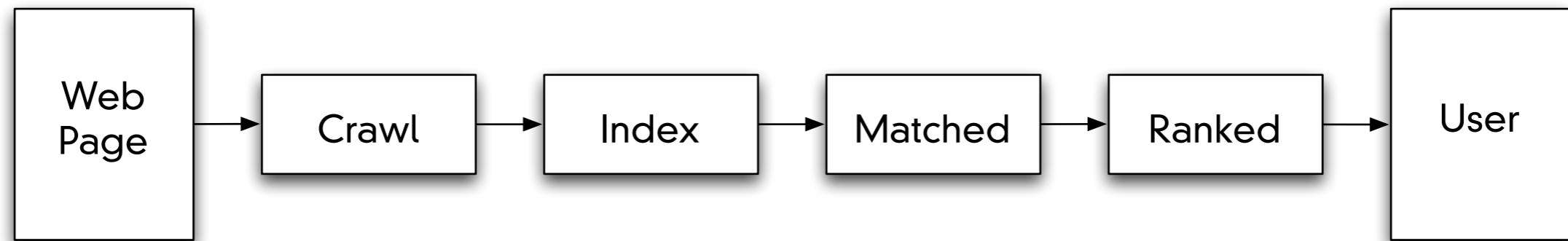
Link-unaware model. If only the head node i knows the existence of the link and the weight value, but the tail node j knows nothing, this is designated as *link-unaware weight-unaware model* (or *link-unaware model* for short). For example, consider a peer evaluation scheme among members of personnel. Each member can choose a limited number of other members.

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



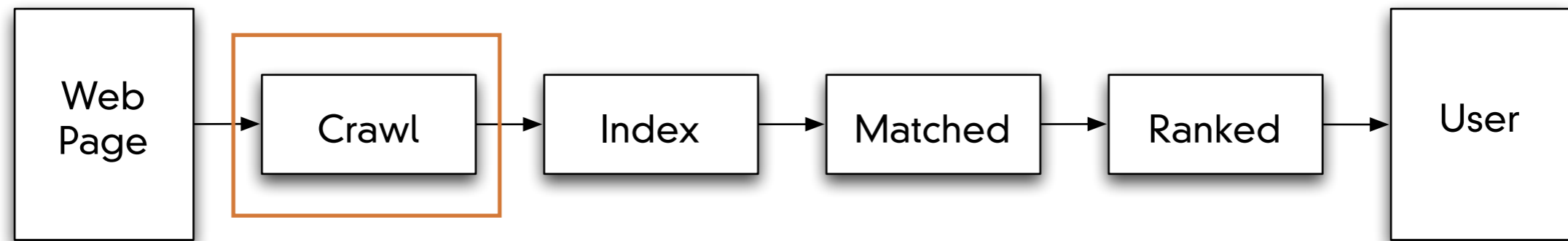
The Impact of Crawl Policy on Web Search Effectiveness



- 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



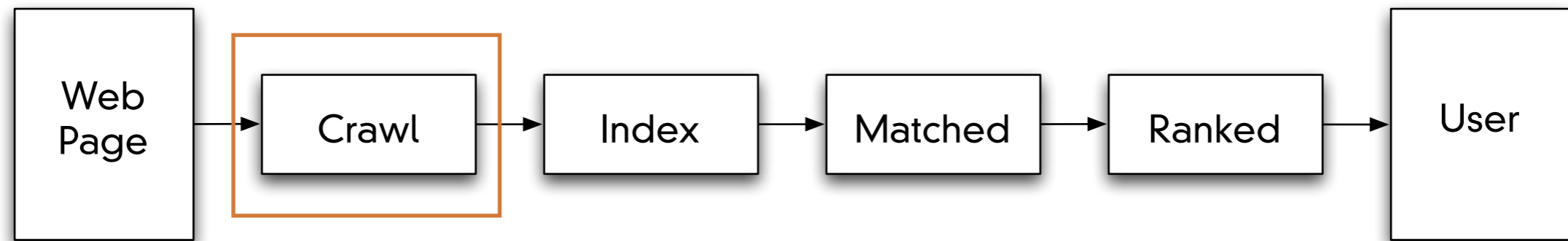
The Impact of Crawl Policy on Web Search Effectiveness



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



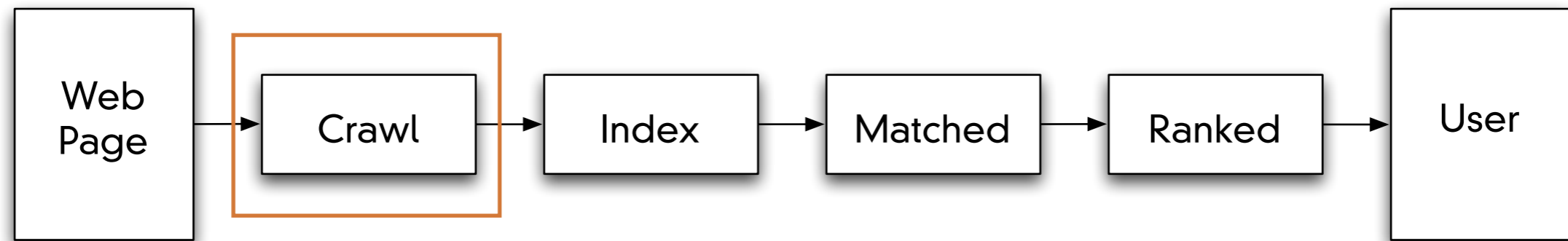
The Impact of Crawl Policy on Web Search Effectiveness



- 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 coefficient**
- 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
 -

Arts

[Movies](#), [Television](#), [Music...](#)

Games

[Video Games](#), [RPGs](#), [Gambling...](#)

Kids and Teens

[Arts](#), [School Time](#), [Teen Life...](#)

Reference

[Maps](#), [Education](#), [Libraries...](#)

Shopping

[Clothing](#), [Food](#), [Gifts...](#)

World

[Català](#), [Dansk](#), [Deutsch](#), [Español](#), [Français](#), [Italiano](#), [日本語](#), [Nederlands](#), [Polski](#), [Русский](#), [Svenska...](#)

Business

[Jobs](#), [Real Estate](#), [Investing...](#)

Health

[Fitness](#), [Medicine](#), [Alternative...](#)

News

[Media](#), [Newspapers](#), [Weather...](#)

Regional

[US](#), [Canada](#), [UK](#), [Europe...](#)

Society

[People](#), [Religion](#), [Issues...](#)

Computers

[Internet](#), [Software](#), [Hardware...](#)

Home

[Family](#), [Consumers](#), [Cooking...](#)

Recreation

[Travel](#), [Food](#), [Outdoors](#), [Humor...](#)

Science

[Biology](#), [Psychology](#), [Physics...](#)

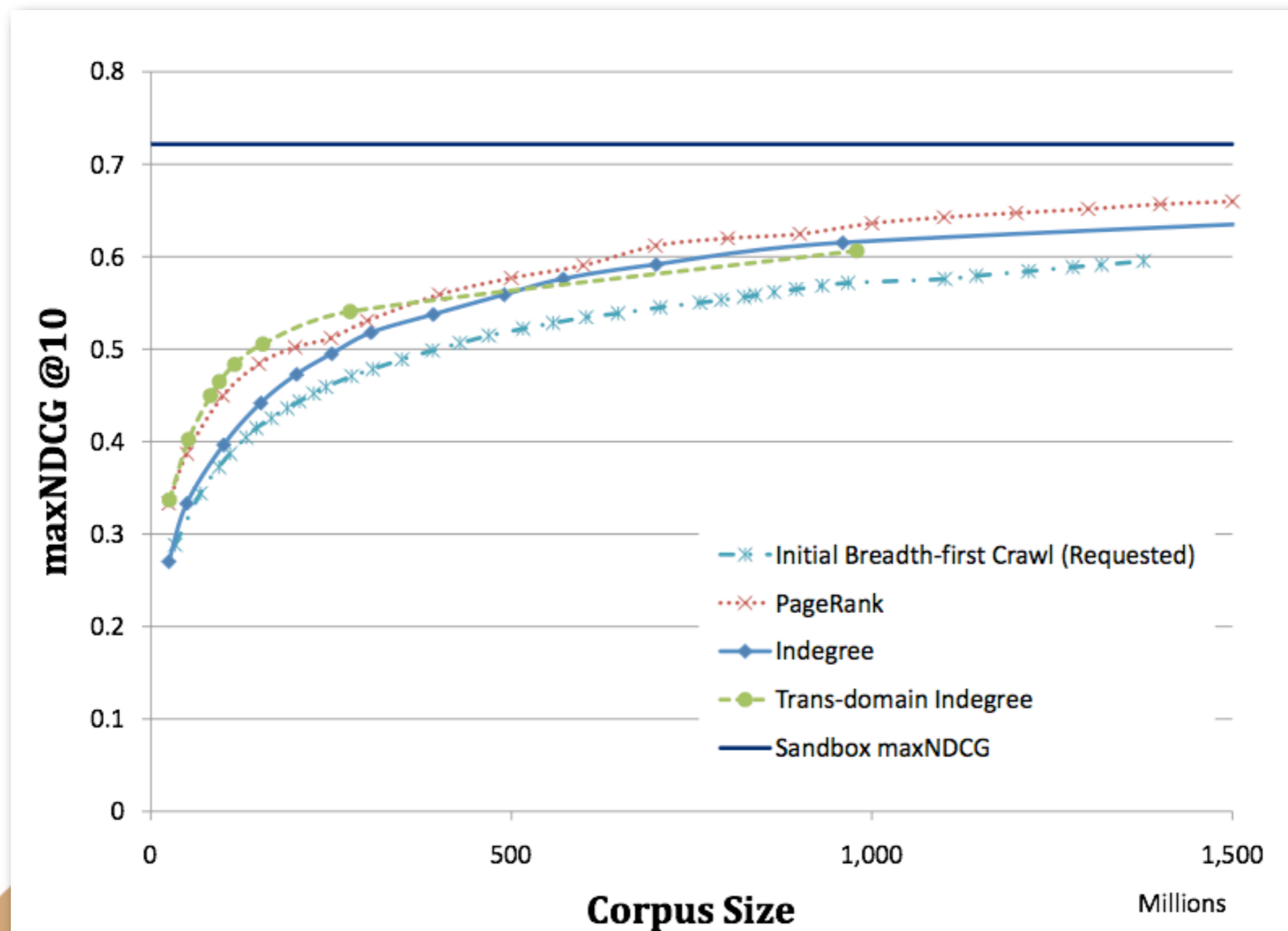
Sports

[Baseball](#), [Soccer](#), [Basketball...](#)



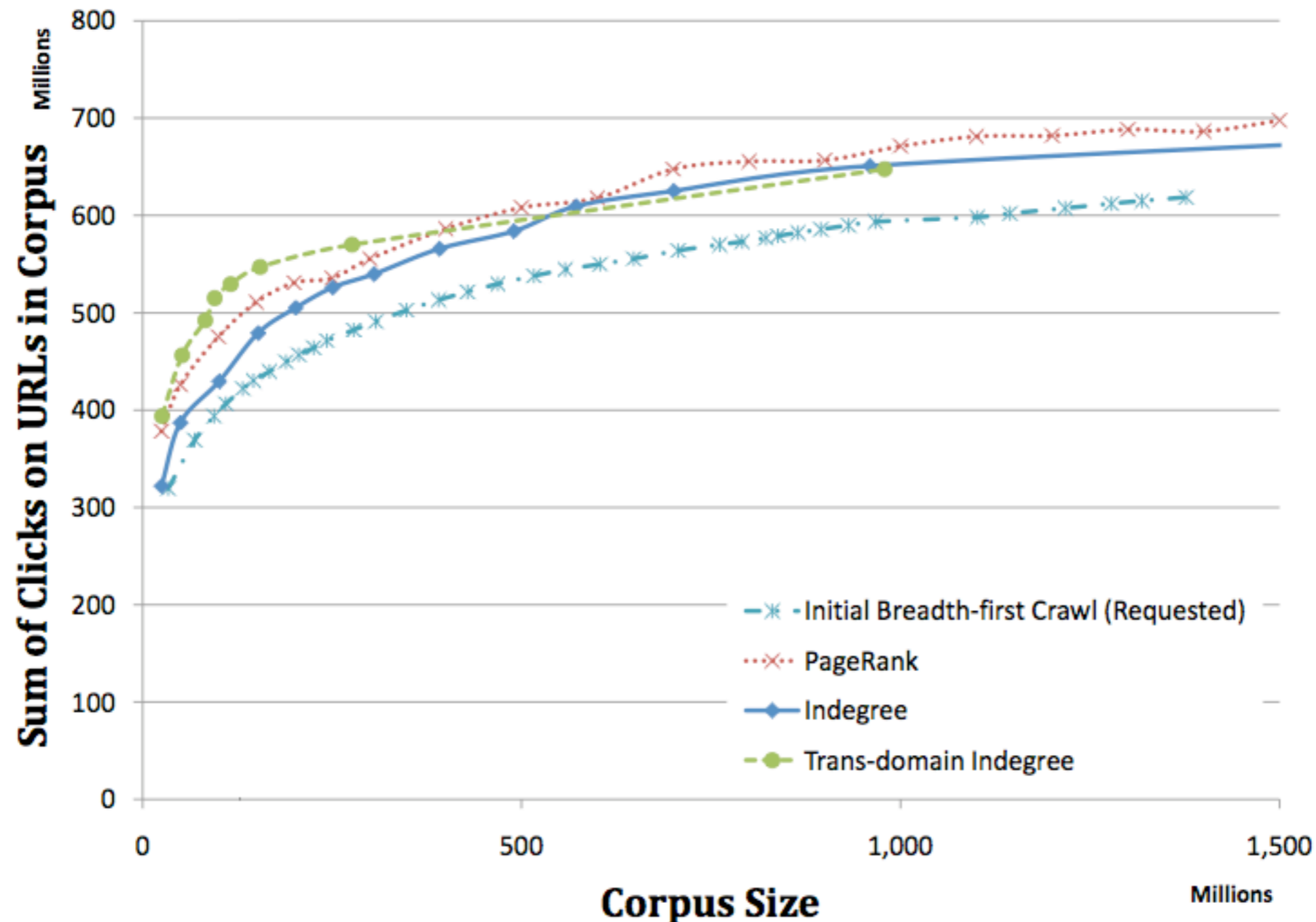
The Impact of Crawl Policy on Web Search Effectiveness

- Experiments - Single Crawl



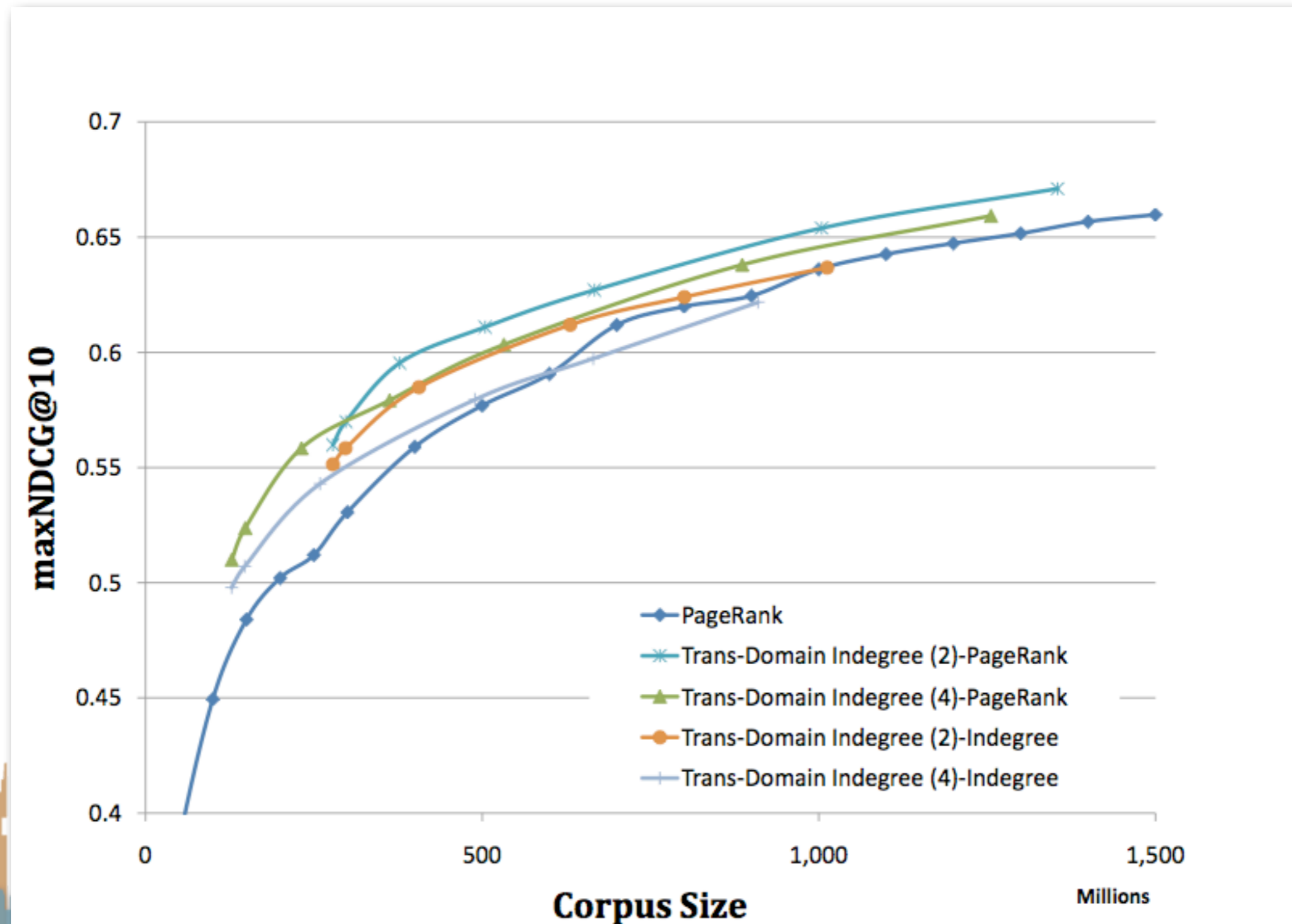
The Impact of Crawl Policy on Web Search Effectiveness

- Experiments - Single Crawl



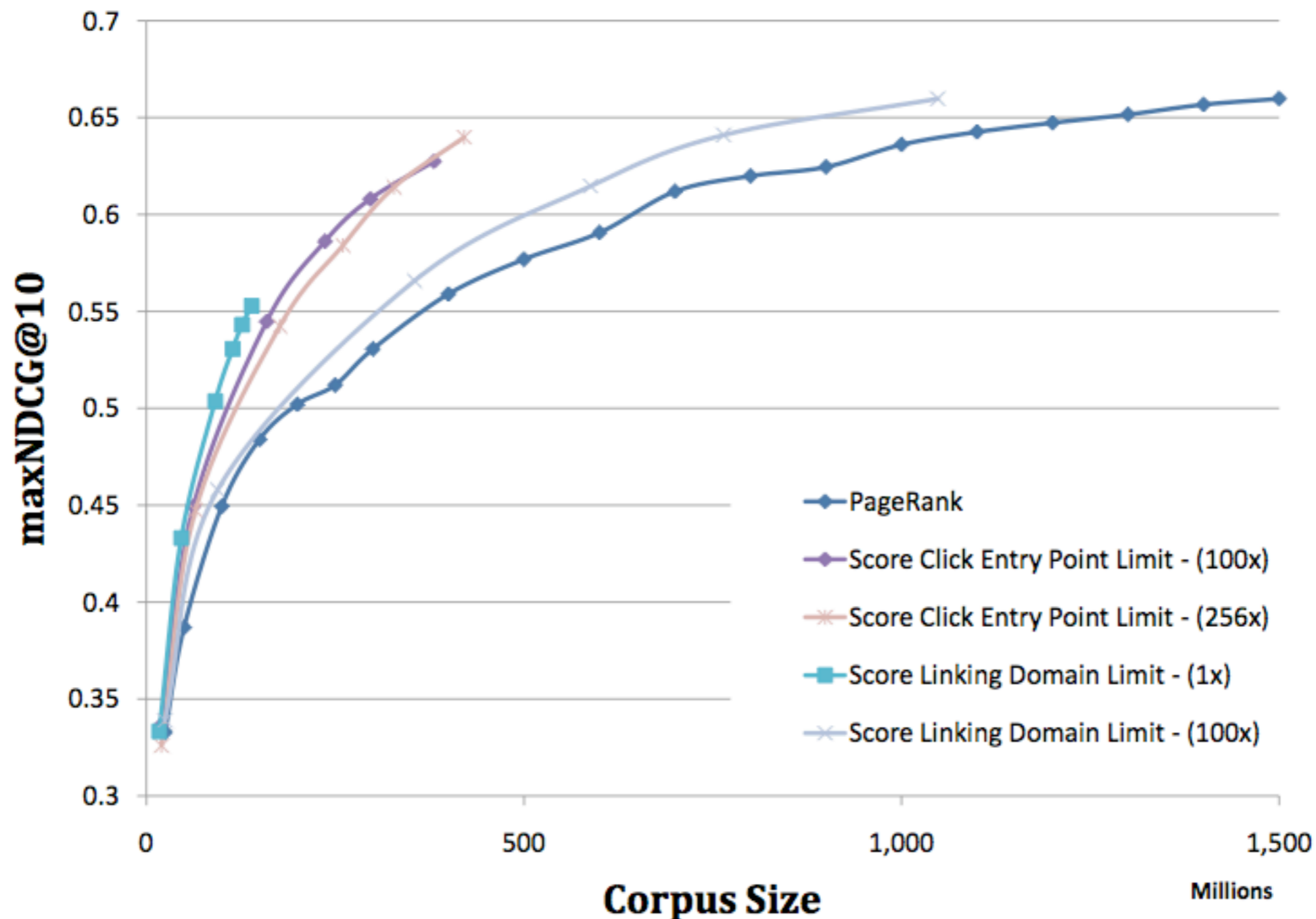
The Impact of Crawl Policy on Web Search Effectiveness

- 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

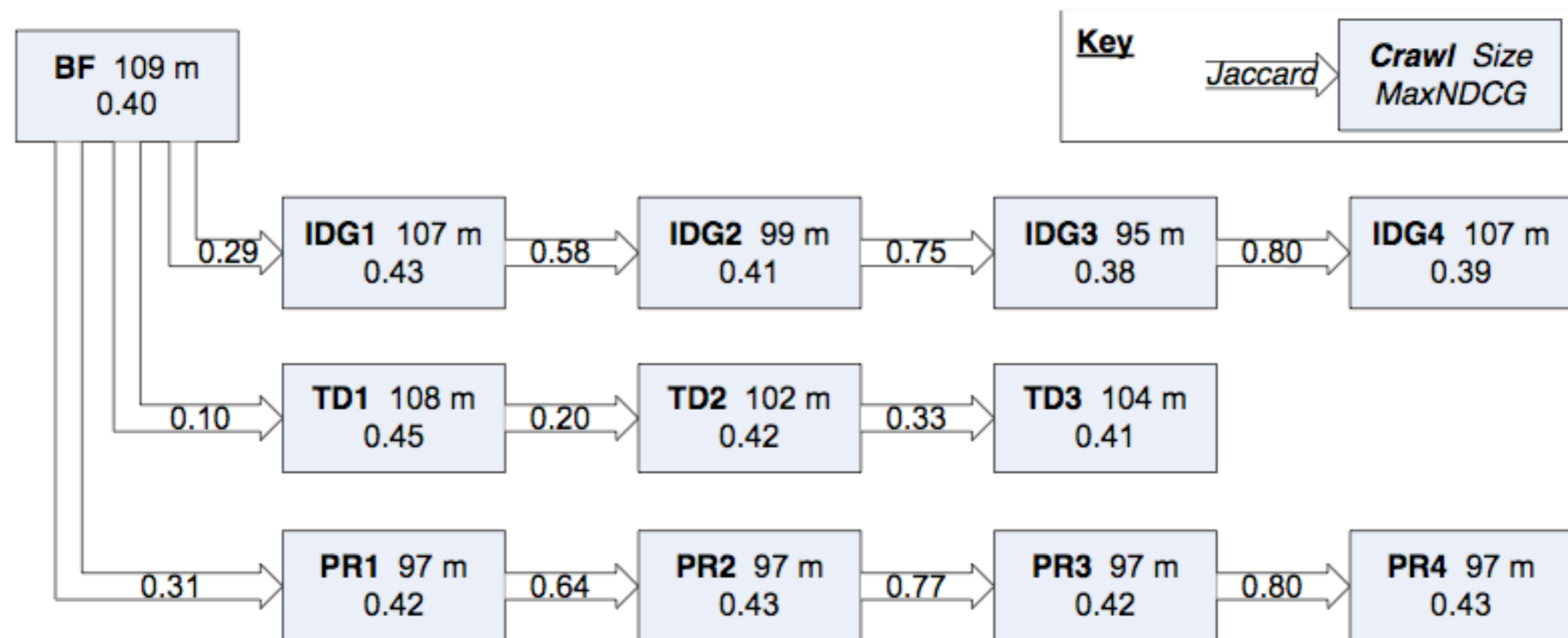


Figure 4: Iterative selection experiments, showing the size and maxNDCG of requested selections, and the Jaccard similarity between iterations.

The Impact of Crawl Policy on Web Search Effectiveness

- Experiments - Iterative Set Similarity

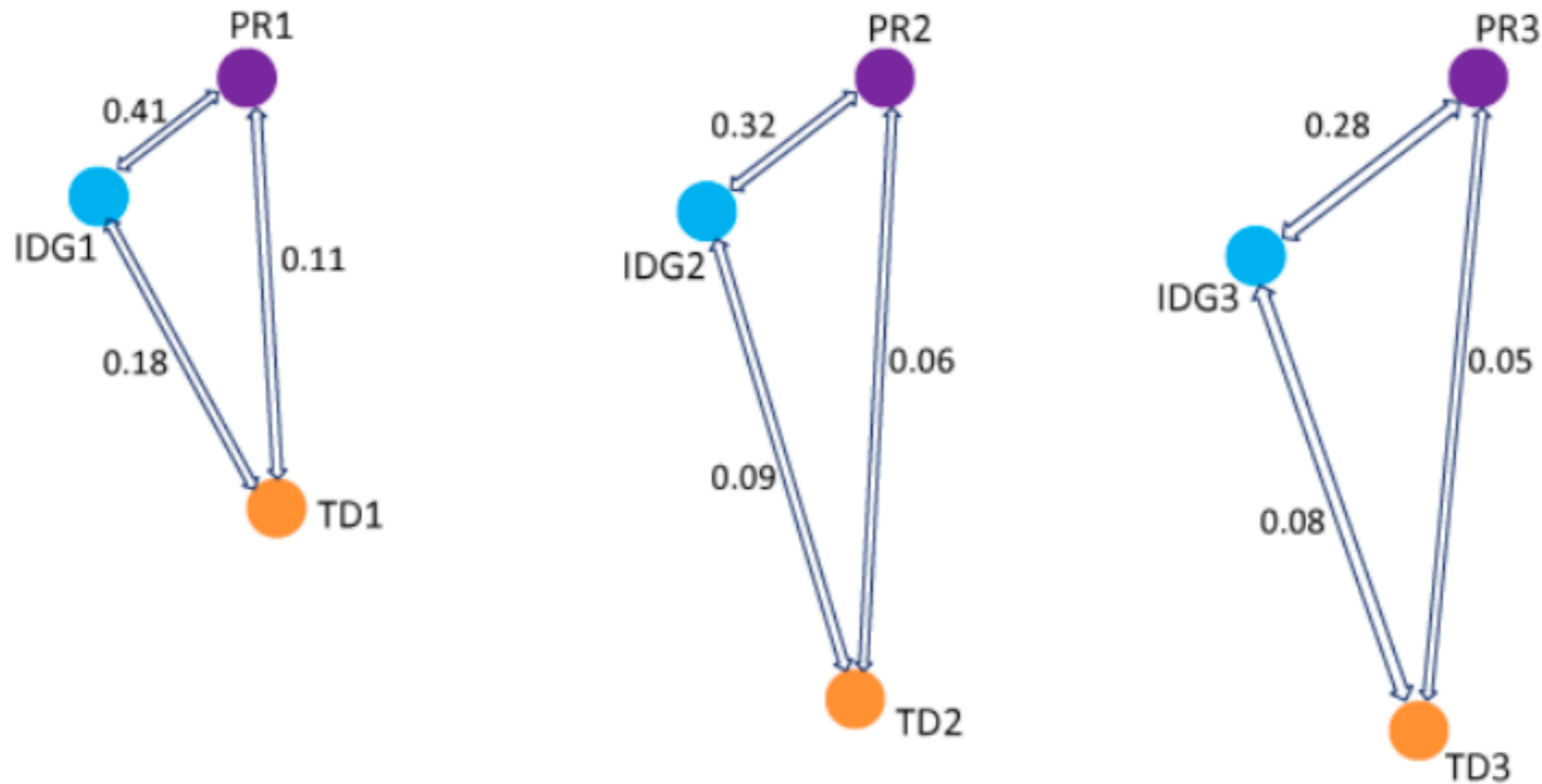
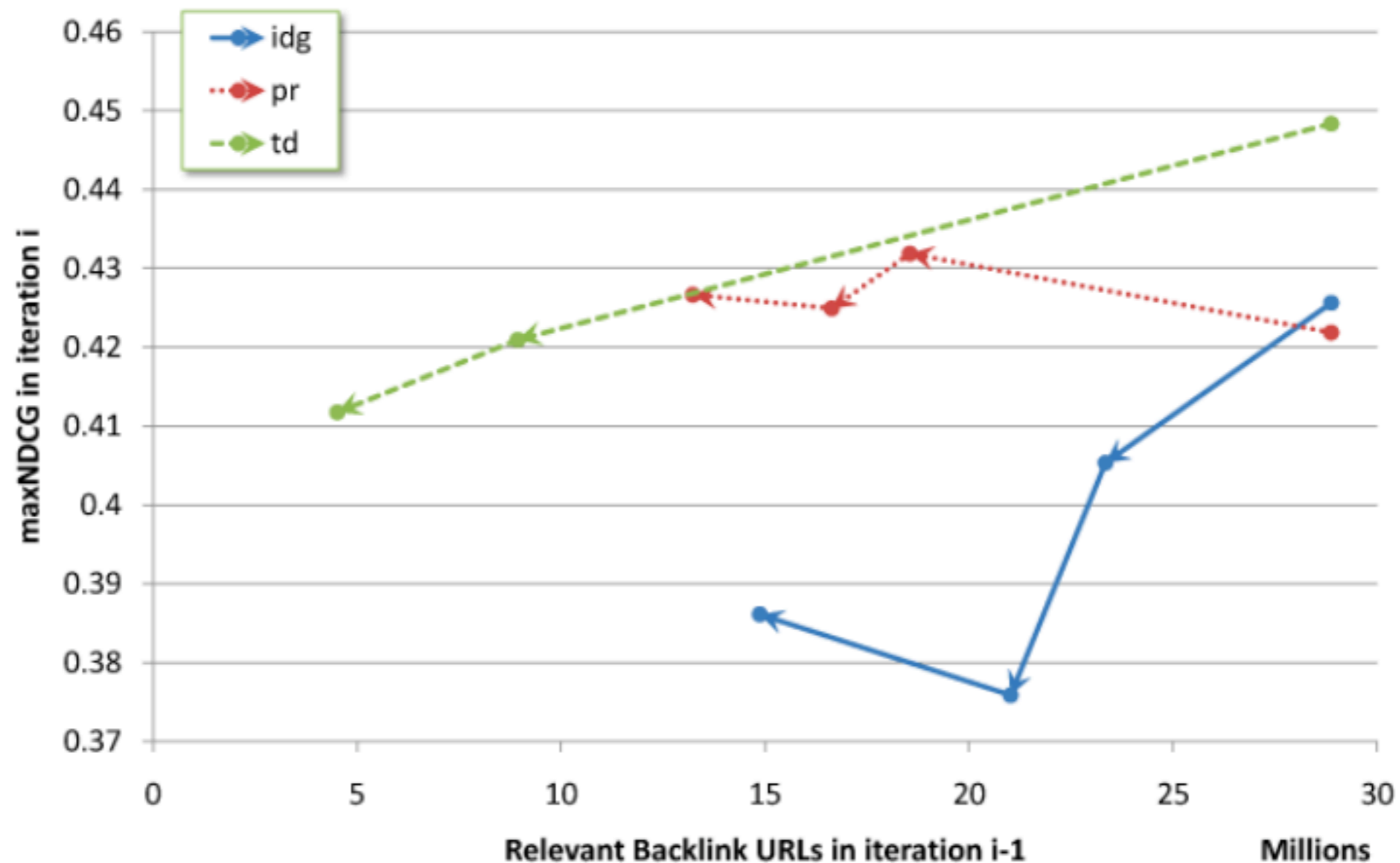


Figure 5: Jaccard similarity between corpora selected by different policies at different iterations.



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

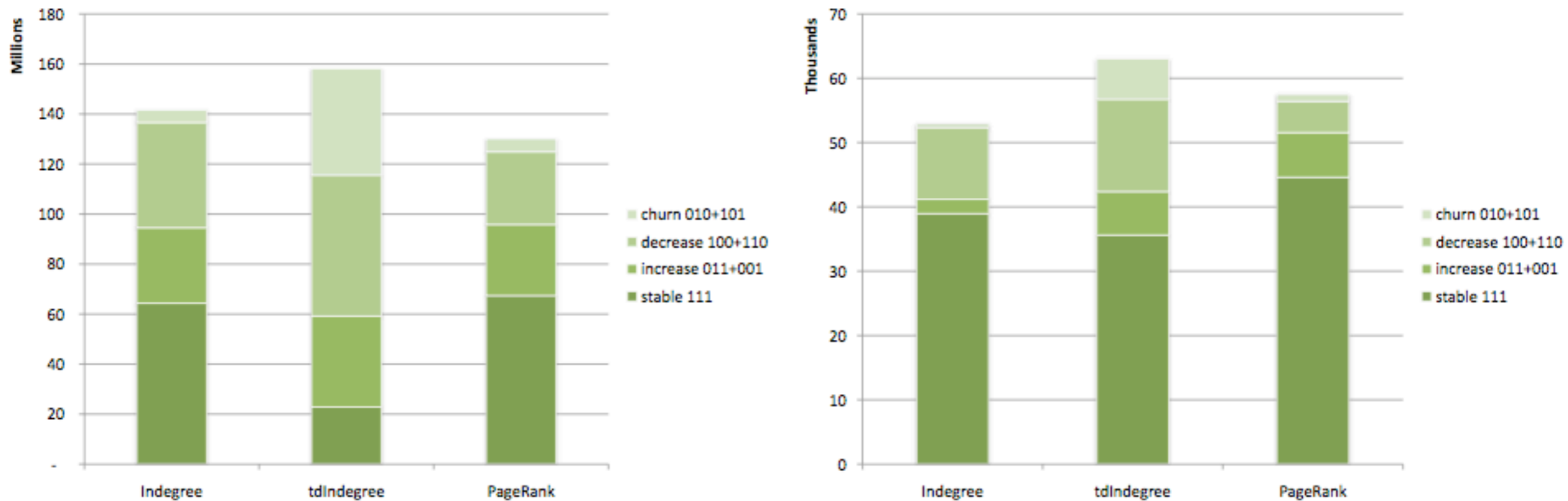


Figure 6: Stability of crawl policies, over three iterations. Left: All URLs. Right: Relevant URLs



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

ABSTRACT

Crawl selection policy has a direct influence on Web search effectiveness, because a useful page that is not selected for crawling will also be absent from search results. Yet there has been little or no work on measuring this effect. We introduce an evaluation framework, based on relevance judgments pooled from multiple search engines, measuring the maximum potential NDCG that is achievable using a particular crawl. This allows us to evaluate different crawl policies and investigate important scenarios like selection stability over multiple iterations. We conduct two sets of crawling experiments at the scale of 1 billion and 100 million pages respectively. These show that crawl selection based on PageRank, indegree and trans-domain indegree all allow better retrieval effectiveness than a simple breadth-first crawl of the same size. PageRank is the most reliable and effective method. Trans-domain indegree can outperform PageRank, but over multiple crawl iterations it is less effective and more unstable. Finally we experiment with combinations of crawl selection methods and per-domain page limits, which yield crawls with greater potential NDCG than PageRank.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms

Measurement, Experimentation

1. INTRODUCTION

A useful Web search result will only be seen by users if it is crawled by the search engine, indexed correctly, found in the index when matched with a query and ranked highly in the search result listing. It only takes one failure in this chain of events for the useful (relevant) result to be lost. If such failures happen often, users will perceive a drop in the quality of search results. Therefore, to optimize user satisfaction, it is important to avoid failure at every stage.

Success at the crawling stage depends on the size of the crawl and the crawl selection policy. For example, the policy of preferring

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
Copyright 2009 ACM 978-1-60558-483-6/09/07 ...\$5.00.

pages with highest PageRank [7] and a size limit of N leads to selecting a set of N high-PageRank pages. When searches are carried out, the quality of search results will sometimes be reduced because pages that would have been relevant and retrievable were not selected for crawling. One way to reduce such failures is to increase the size N of the crawl. Another approach is to improve the selection policy.

Although well-known methods exist for evaluating search relevance, such as NDCG [13], we are not aware of any published experiments that compare the relevance achievable by different crawl policies. Acting as a barrier to experimentation are the large communication and computational costs of conducting multiple crawls, creating multiple indices and processing queries. Our framework ameliorates this via a crawl sandbox and an evaluation metric that only requires the set of selected URLs. The sandbox is simply a cache, to avoid crawling URLs more than once if selected by multiple policies or iterations. The metric, maxNDCG, is the best potential NDCG that could be achieved based on the presence or absence of relevant pages in a crawl. maxNDCG is proportional to NDCG but may be calculated without indexing and retrieval. It may even be calculated for a selected set of pages without attempting to crawl them, estimating the NDCG that would be achievable by a perfect ranker if all selected pages were successfully crawled.

These efficiency techniques allow us to run a large number of experiments comparing crawl policies. We focus on policy selection based on the link graph of a previous crawl. This is a common scenario, allowing an engine to shift its focus towards pages that are preferred according to some link-based metric (such as PageRank) but not yet included in the crawl.

In Section 2, we discuss different aspects involved in the selection of crawling methods. We provide motivation for our experimental setup, and where appropriate, we provide references to other work. We then present experiments in Section 3 and Section 4.

2. CRAWLING AND EVALUATION

Search engines are the primary discovery mechanism on the Web, and the Web has an effectively infinite number of pages that might be indexed. A search engine must select a subset of pages to make the best use of its resources. Search engines use crawlers to download pages and extract links. It is important to select which indices are built. Starting from a set of seed URLs, which indices are built. Starting from a set of seed URLs, which indices are built. Starting from a set of seed URLs, which indices are built.

After an initial crawl, there is a corpus, since pages are continually deleted [4]. One option would be to delete completely throwing away

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

ABSTRACT

Link analysis methods have been used successfully for knowledge discovery from the link structure of mutually linking entities. Existing link analysis methods have been inherently designed based on the fact that the entire link structure of the target graph is observable such as public web documents; however, link information in graphs in the real world, such as human relationship or economic activities, is rarely open to public. If link analysis can be performed using graphs with private links in a privacy-preserving way, it enables us to rank entities connected with private ties, such as people, organizations, or business transactions. In this paper, we present a secure link analysis for graphs with private links by means of cryptographic protocols. Our solutions are designed as privacy-preserving expansions of well-known link analysis methods, PageRank and HITS. The outcomes of our protocols are completely equivalent to those of PageRank and HITS. Furthermore, our protocols theoretically guarantee that the private link information possessed by each node is not revealed to other nodes.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous
Algorithms

General Terms

Algorithms

Keywords

link analysis, privacy, ranking, HITS, PageRank

1. INTRODUCTION

Link-based analysis has been developed in the form of algorithms that discover useful information from the link structure of mutually linking entities. In particular, HITS [7] and PageRank [9] have been successfully used for the ranking of hyperlinked web documents. These link analysis methods were originally designed for the analysis of web documents; however, these can be readily applied to mutually linking entities, such as referenced academic papers, protein-protein interactions, and so on.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
Copyright 2009 ACM 978-1-60558-483-6/09/07 ...\$5.00.

In general, link analysis methods take the entire link structure as its input. Indeed, for the computation of Google's PageRank, the linking structures of web documents are collected by crawling agents which actually wander around public web documents. The same holds for citation graphs of academic papers or interaction graphs of protein networks. As shown, existing link analysis methods have inherently been designed based on the fact that the entire link structure of the target graph is observable; however, link information in the real world, such as human relationships or economic activities, is rarely open to public.

In this paper, we present link analysis solutions for graphs of privately connected entities. Let there be a directed weighted graph $G = (V, E, W)$ where V is a set of vertices, E is a set of edges, and W is a weight matrix. Throughout this paper, we assume that the set of vertices corresponds to a collection of distributed nodes where the computational power of each node is polynomial. Edges correspond to links between nodes; weights of edges correspond to weights of these links. Let there be a link of node i pointing to node j . In our setting, we assume that link e_{ij} and weight of the link w_{ij} are not desired to be known by nodes other than node i and node j . Furthermore, we design our link analysis solutions based on the three privacy models of graphs described as below:

Weight-aware model. If both the head node i and the tail node j know the existence of the link and the weight value, this is designated as *weight-aware link-aware model* (or *weight-aware model* for short). For example, consider commercial relationships among enterprises. Each enterprise may conduct business transactions with the other enterprises. Let the i th enterprise purchase some products from the j th enterprise. This transaction corresponds to link e_{ij} and the transaction value corresponds to weight w_{ij} . In this case, both the i th and j th enterprise are aware of the existence of this link and know the weight value, but enterprises other than i and j do not know the existence of this transaction and the transaction value.

Link-aware model. If the head node i and the tail node j know the existence of the link, but the weight value is only known by the head node i , this is designated as *link-aware weight-unaware model* (or *link-aware model* for short). For example, consider call logs of cell-phones. Let caller i make a phone call to receiver j . This call corresponds to link e_{ij} and the probability that i makes a phone call to j corresponds to the weight w_{ij} of e_{ij} . In this case, both caller i and receiver j are aware of the existence of the link, but the caller probability w_{ij} are known only by caller i .

Link-unaware model. If only the head node i knows the existence of the link and the weight value, but the tail node j knows nothing, this is designated as *link-unaware weight-unaware model* (or *link-unaware model* for short). For example, consider a peer evaluation scheme among members of personnel. Each member can choose a limited number of other members.