course:

Searching the Web and Multimedia Databases (BI-VWM)

SS2011/12

lecture 7:

Searching the deep web

based on **Exploration of Deep Web Repositories**

by **Nan Zhang**, The George Washington Uni., **Gautam Das**, Uni. of Texas, Arlington

tutorial given at **VLDB 2011** conference (see the uncut version here)

doc. RNDr. Tomáš Skopal, Ph.D.

Department of Software Engineering, Faculty of Mathematics and Physics, Charles University in Prague

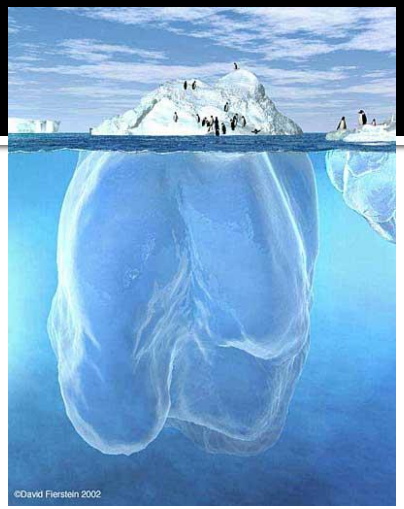
Department of Software Engineering, Faculty of Information Technology, Czech Technical University in Prague

https://edux.fit.cvut.cz/courses/BI-VWM/

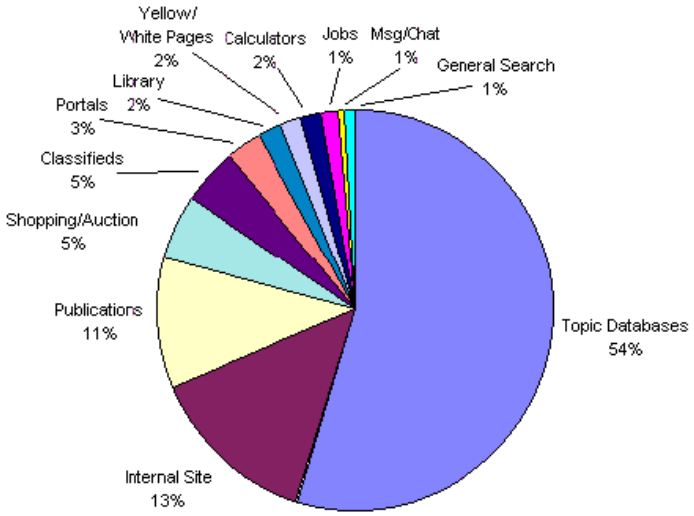
Outline

- Introduction
- Resource discovery and interface understanding
- Crawling
- Sampling
- *Data Analytics (not covered in this cut version)*

The Deep Web

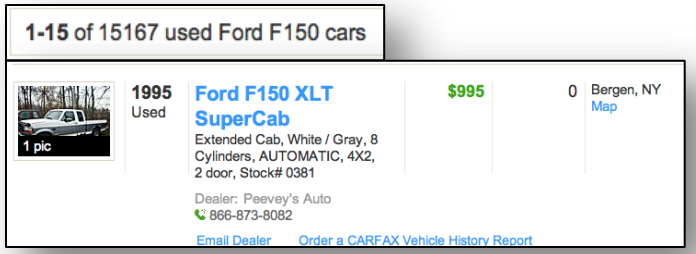


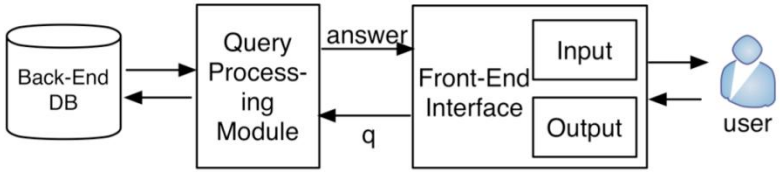
- Deep Web vs Surface Web
  - Dynamic contents, unlinked pages, private web, contextual web, etc
  - Estimated size:
    91,850 vs 167 tera bytes[1], hundreds or thousands of times larger than the surface web[2]

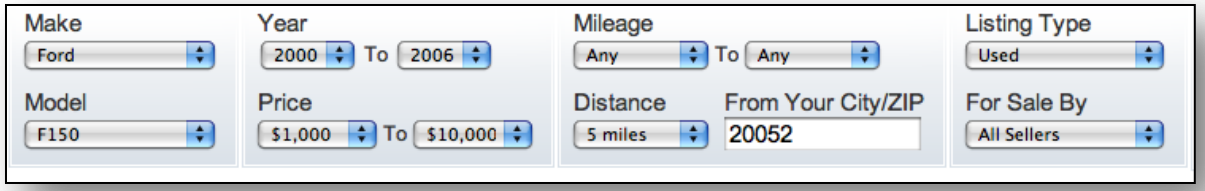


[1] SIMS, UC Berkeley, How much information? 2003
[2] Bright Planet, Deep Web FAQs, 2010, http://www.brightplanet.com/the-deep-web/

Hidden Web Repositories







Hidden Repository | Web
Owner | User

Deep Web Repository: Example I

Enterprise Search Engine's Corpus

Unstructured data | Keyword search | Top-k

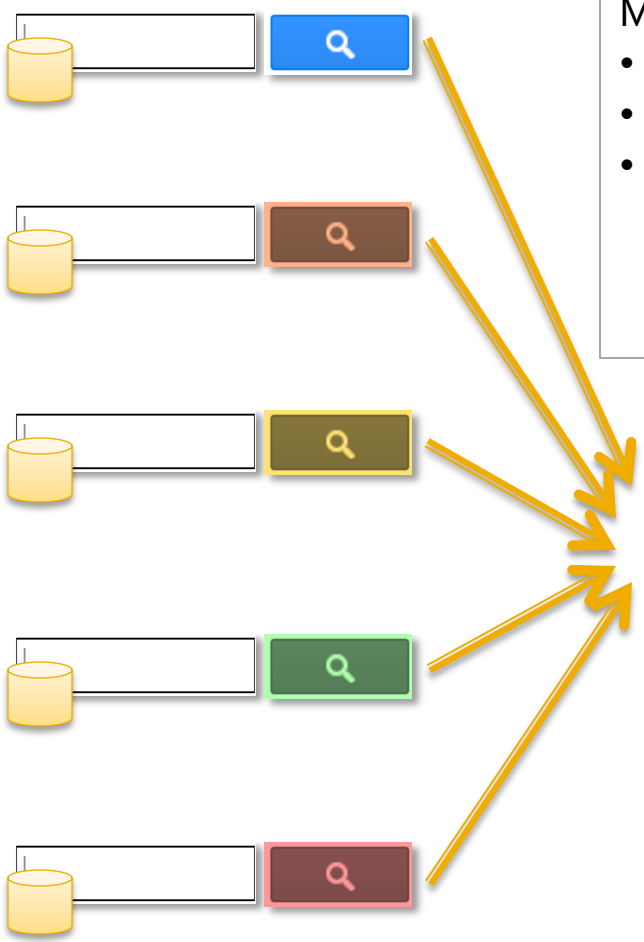
CDC - **Asthma** and Allergies - Prevention of Occupational **...**
ASTHMA AND ALLERGIES. Prevention of Occupational **Asthma**: Introduction. **...** Smith AM, Bernstein DI. Management of work-related **asthma**. **...**
www.cdc.gov/niosh/topics/asthma/OccAsthmaPrevention.html
[More results from www.cdc.gov/niosh/topics/asthma]

Lower Airway Rhinovirus Burden and the Seasonal Risk of **Asthma** Exacerbation.
Denlinger LC, Sorkness RL, Lee WM, Evans M, Wolff M, Mathur S, Crisafi G, Gaworski K, Pappas
Am J Respir Crit Care Med. 2011 Aug 4. [Epub ahead of print]
PMID: 21816938 [PubMed - as supplied by publisher]
Related citations

First Aid/CPR/AED - Professional Rescuers
... one- and two-rescuer); AED; Optional training in use of epinephrine auto-injectors and **asthma** inhalers available. Course ...

- Brand Name: Foradil Aerolizer
  Generic Name: Formoterol Fumarate Inhalation Powder
- Brand Name: Qvar
  Generic Name: Beclomethasone Dipropionate HFA

Exploration: Example I



Metasearch engine

- Discovers deep web repositories of a given topic
- Integrate query answers from multiple repositories
- For result re-organization, evaluate the quality of each repository through analytics
  - e.g., how large is the repository?
  - e.g., average length of documents of a given topic

Treatment info

Foradil Aerolizer
e: Formoterol Fumar

Qvar
e: Beclomethasone

Asthma Treatment
AsthmaTreatmentOpt
More-Get Info.

Disease info

CDC - **Asthma** and Allergies - Prevention of Occupational ...
ASTHMA AND ALLERGIES. Prevention of Occupational **Asthma**: Management of work-related **asthma**. ...
www.cdc.gov/niosh/topics/asthma/OccAsthmaPrevention.html
[More results from www.cdc.gov/niosh/topics/asthma]

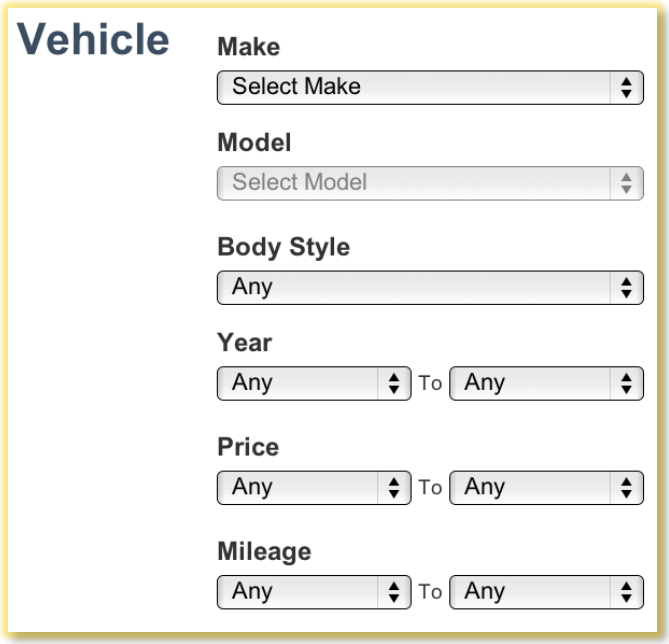
Lower Airway Rhinovirus Burden and the Seasonal Risk of **Asthma**
Denlinger LC, Sorkness RL, Lee WM, Evans M, Wolff M, Mathur S,
Am J Respir Crit Care Med. 2011 Aug 4. [Epub ahead of print]
PMID: 21816938 [PubMed - as supplied by publisher]
Related citations

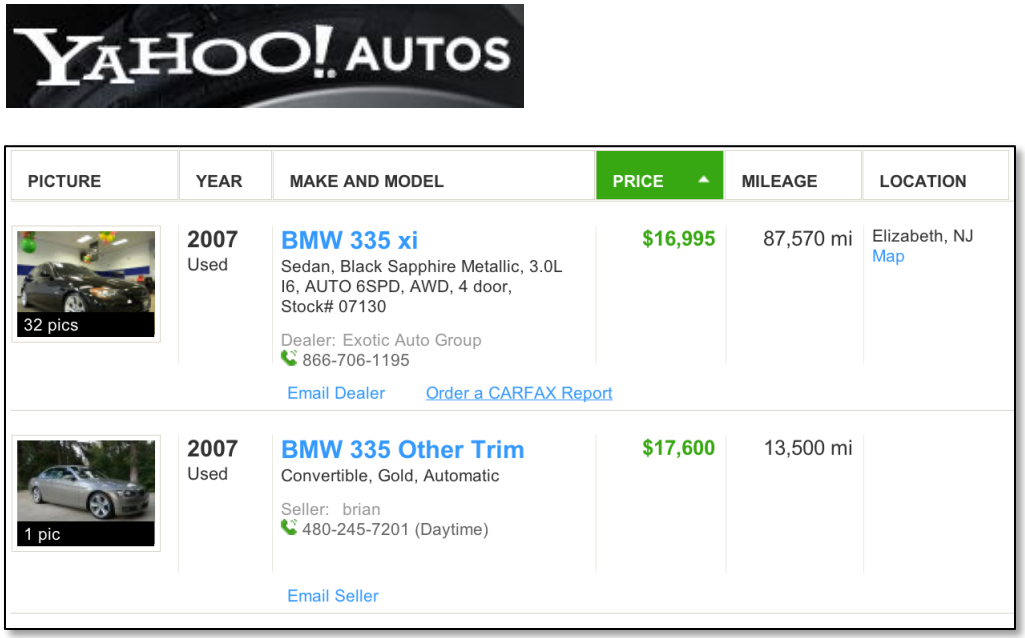
dogpile TOPSY YIPPY! YAHOO! WebCrawler®

Example II

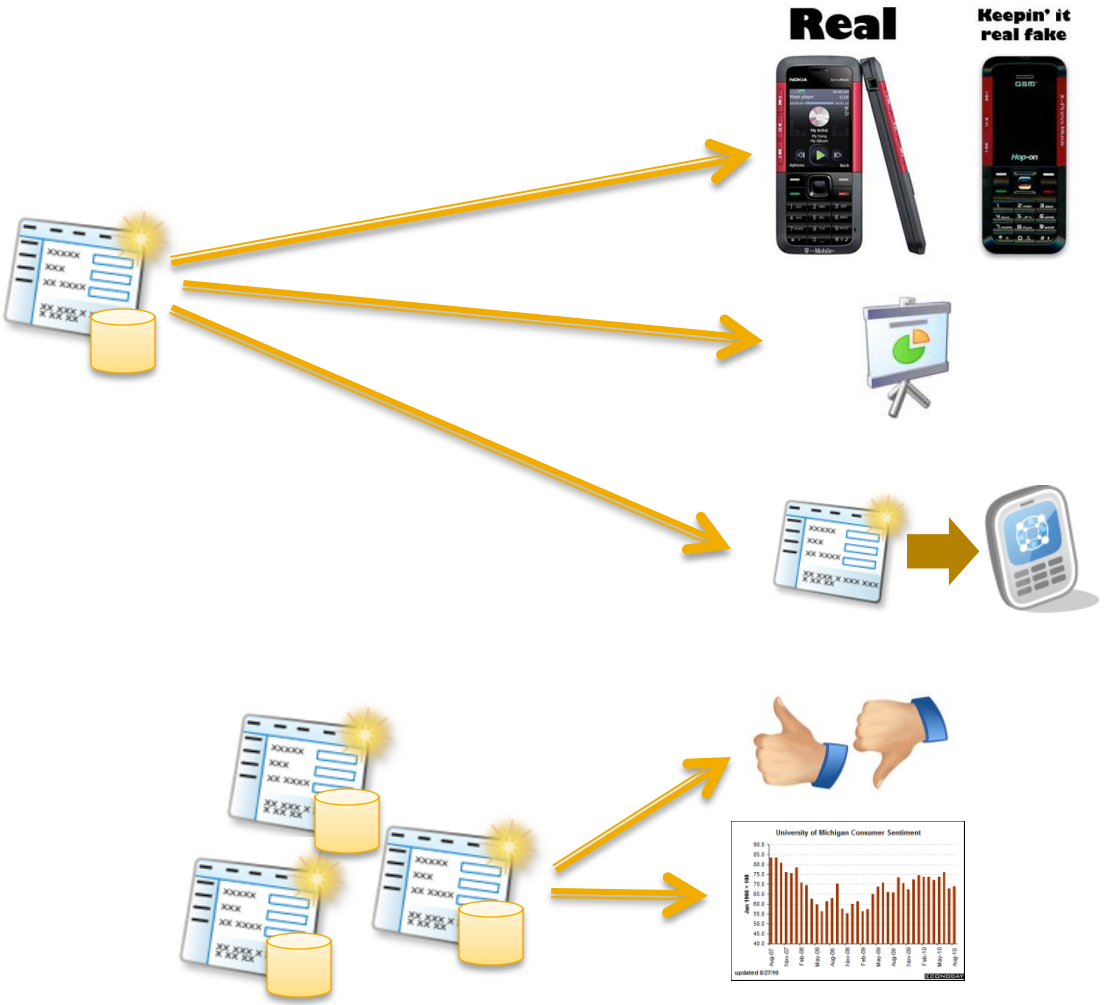
Yahoo! Auto, other online e-commerce websites

Structured data | Form-like search | Top-1500





Exploration: Example II



Third-party services for an individual repository

- Find fake products
- Price distribution
- Construction of a universal mobile interface

Third-party services for multiple repositories

- Repository comparison
- Consumer behavior analysis

Main Tasks

- Resource discovery
- Data integration
- Single-/Cross- site analytics

KAYAK

TERAPEAK

pricegrabber

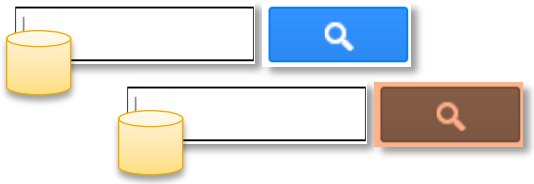
Summary of Main Tasks/Obstacles

- Find where the data are
  - Resource discovery: find URLs of deep web repositories
  - Required by: Metasearch engine, shopping website comparison, consumer behavior modeling, etc.
- Understand the web interface
  - Required by almost all applications.
- Explore the underlying data
  - crawling, sampling, and analytics
  - Required by: Metasearch engine, keep it real fake, price prediction, universal mobile interface, shopping website comparison, consumer behavior modeling, market penetration analysis, social page evaluation and optimization, etc.

Covered by many recent tutorials [Weikum and Theobald PODS 10, Chiticariu et al SIGMOD 10, Dong and Nauman VLDB 09, Franklin, Halevy and Maier VLDB 08]

Demoed by research prototypes and product systems

DBLife WEBTABLES TEXTRUNNER Arnetminer Cazoodle PATENTSSEARCHER



Focus of This Lecture

- Brief Overview of:
  - Resource discovery
  - Interface understanding
  - i.e., where to, and how to issue a search query to a deep web repository?
- Our focus: Data crawling, sampling, and analytics

Which individual search and/or browsing requests should a third-party explorer issue to the the web interface of a given deep web repository, in order to enable efficient crawling, sampling, and data analytics?

Outline

- Introduction
- Resource discovery and interface understanding
- Crawling
- Sampling

Resource Discovery

- Objective: discover resources of "interest"
  - Task 1: is an URL of interest?
    - Criteria A: is a deep web repository
    - Criteria B: belongs to a given topic
  - Task 2: Find all interesting URLs
- Task 1, Criteria A
  - Transactional page search [LKV+06]
    - Pattern identification – e.g., "Enter keywords", form identification
    - Synonym expansion – e.g., "Search" + "Go" + "Find it"
- Task 1, Criteria B:
  - Learn by example
- Task 2
  - Topic distillation based on a search engine
    - e.g., "used car search", "car * search"
    - Alone not suffice for resource discovery [Cha99]
  - Focused/Topical "Crawling"
    - Priority queue ordered by importance score
    - Leveraging locality
    - Often irrelevant pages could lead to relevant ones
      - Reinforcement learning, etc.

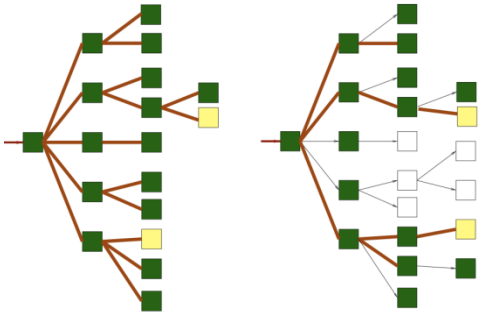


Figure from [DCL+00]

[DCL+00] M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles, and M. Gori, "Focused crawling using context graphs", VLDB, 2000.

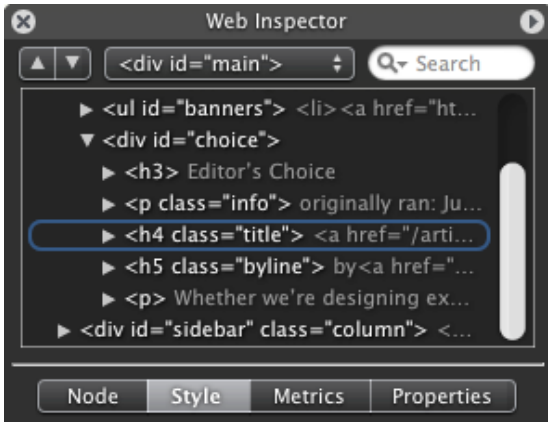
[LKV+06]Y. Li, R. Krishnamurthy, S. Vaithyanathan, and H. V. Jagadish, "Getting Work Done on the Web: Supporting Transactional Queries", SIGIR, 2006.

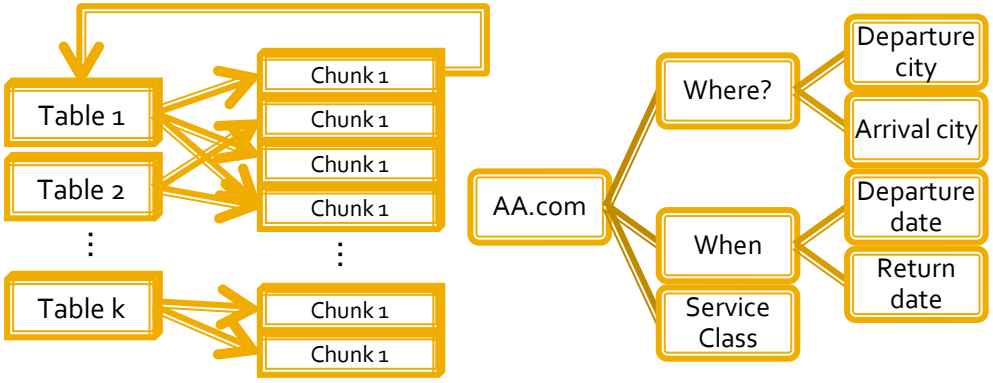
[Cha99] S. Chakrabarti, "Recent results in automatic Web resource discovery", ACM Computing Surveys, vol. 31, 1999.

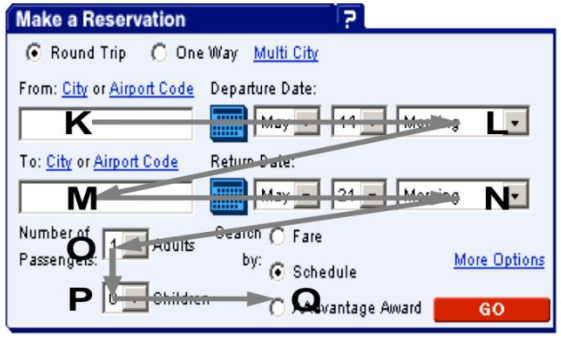
Interface Understanding

Modeling Web Interface

- Generally easy for keyword search interface, but can be extremely challenging for others (e.g., form-like search, graph-browsing)
- What to understand?
  - Structure of a web interface
- Modeling language
  - Flat model e.g., [KBG+01]
  - Hierarchical model e.g., [ZHC04, DKY+09]
- Input information
  - HTML Tags e.g., [KBG+01]
  - Visual layout of an interface e.g., [DKY+09]





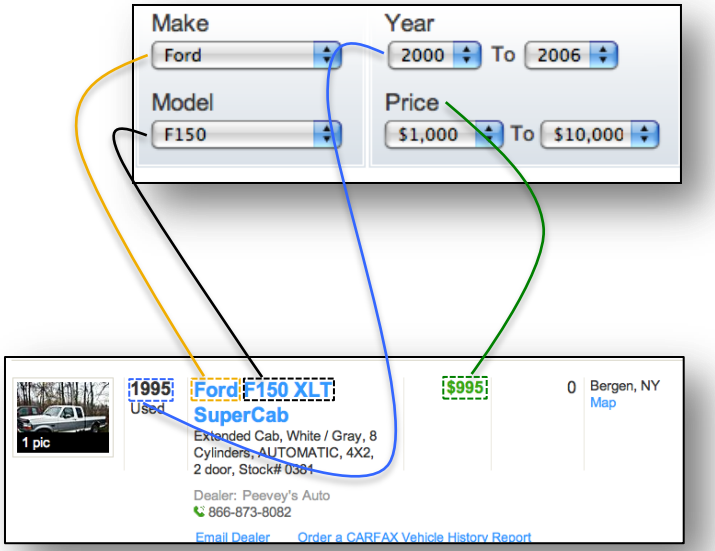


[KBG+01] O. Kaljuvee, O. Buyukkokten, H. Garcia-Molina, and A. Paepcke, "Efficient Web Form Entry on PDAs", WWW 2001.
[ZHC04] Z. Zhang, B. He, and K. C.-C. Chang, "Understanding Web Query Interfaces: Best-Effort Parsing with Hidden Syntax", SIGMOD 2004
[DKY+09] E. C. Dragut, T. Kabisch, C. Yu, and U. Leser, "A Hierarchical Approach to Model Web Query Interfaces for Web Source Integration", VLDB, 2009.

Interface Understanding

Schema Matching

- What to understand?
  - Attributes corresponding to input/output controls on an interface
- Modeling language
  - Map schema of an interface to a mediated schema (with well understood attribute semantics)
- Key Input Information
  - Data/attribute correlation [SDH08, CHW+08]
  - Human feedback [CVD+09]
  - Auxiliary sources [CMH08]



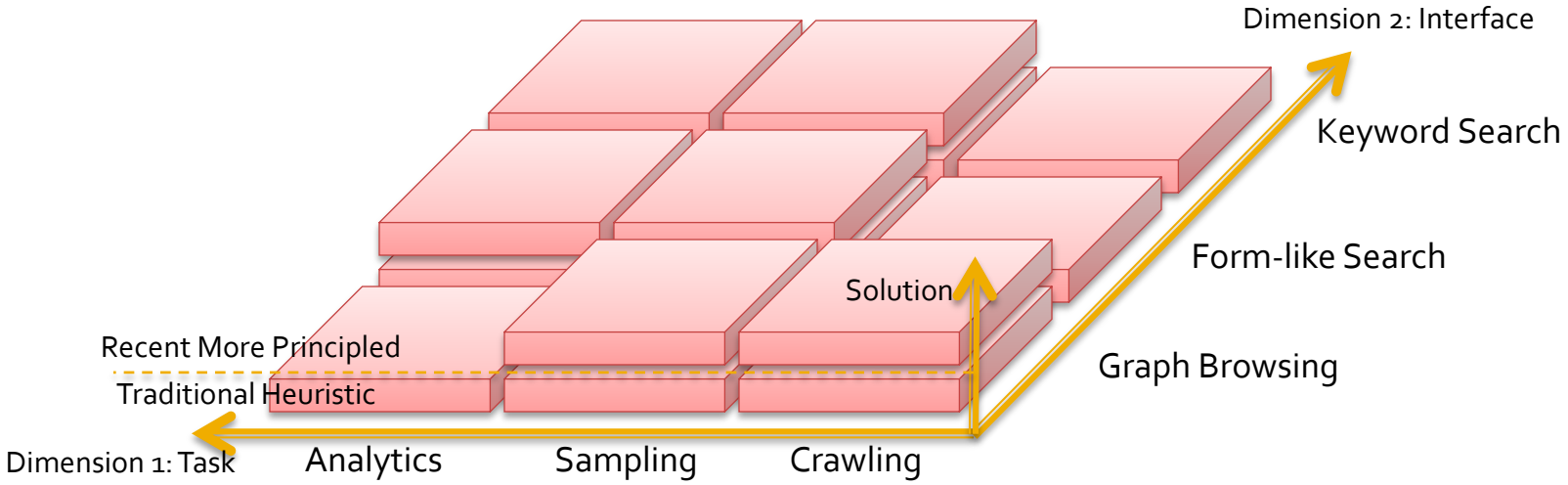
[CHW+08] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang, "WebTables: exploring the power of tables on the web", VLDB, 2008.

[SDH08] A. D. Sarma, X. Dong, and A. Halevy, "Bootstrapping Pay-As-You-Go Data Integration Systems", SIGMOD, 2008.

[CVD+09] X. Chai, B.-Q. Vuong, A. Doan, and J. F. Naughton, "Efficiently Incorporating User Feedback into Information Extraction and Integration Programs", SIGMOD, 2009.

[CMH08] M. J. Cafarella, J. Madhavan, and A. Halevy, "Web-Scale Extraction of Structured Data", SIGMOD Record, vol. 37, 2008.

Problem Space and Solution Space



Problem Space

Solution Space

Around 2000

~ 2005 - now

Traditional Heuristic Approaches

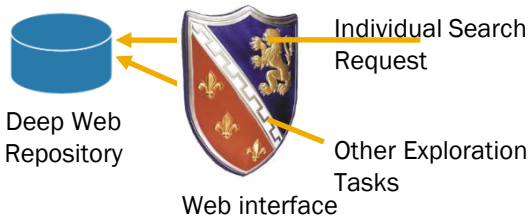
- e.g., seed-query based bootstrapping for crawling
- e.g., query sampling for repository sampling
- No guarantee on query cost, accuracy, etc.

Recent Approaches with Theoretical Guarantees

- e.g., performance-bounded crawlers
- e.g., unbiased samplers and aggregate estimators
- Techniques built upon sampling theory, etc.

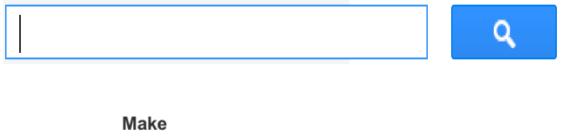
Dimension 1. Task

- Crawling
  - Objective: download as many elements of interest (e.g., documents, tuples, metadata such as domain values) from the repository as possible.
  - Applications: building web archives, private directors, etc.
- Sampling
  - Draw sample elements from a repository according to a pre-determined distribution (e.g., uniform distribution for simple random sampling)
  - Why? Because crawling is often impractical for very large repositories because of practical limitations on the number of web accesses.
  - Collected sample can be later used for analytical processing, mining, etc.
  - Applications: Search-engine quality evaluation for meta-search-engines, price distribution, etc.
- Data Analytics
  - Directly support online analytics over the repository
  - Key Task: efficiently answer aggregate queries (COUNT, SUM, MIN, MAX, etc.)
  - Overlap with sampling, but a key difference on the tradeoff of **versatility** vs. **efficiency**.
  - Applications: consumer behavior analysis, etc.



Dimension 2. Interface

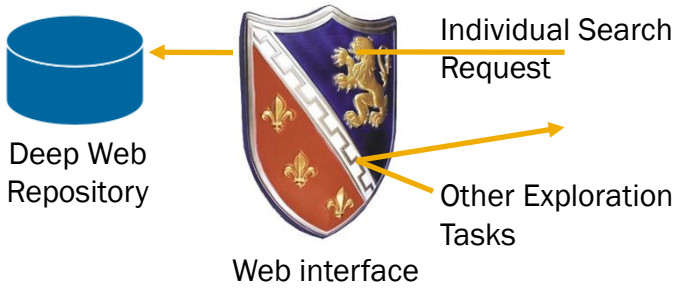
- Keyword-based search
  - Users specify one or a few keywords
  - Common for both structured and unstructured data
  - e.g., Google, Bing, Amazon.
- Form-like search
  - Users specify desired values for one or a few attributes
  - Common for structured data
  - e.g., Yahoo! Autos, AA.com, NSF Award Search.
  - A similar interface: hierarchical browsing
- Graph Browsing
  - A user can observe certain edges and follow through them to access other users' profiles.
  - Common for online social networks
  - e.g., Twitter, Facebook, etc.
- A Combination of Multiple Interfaces
  - e.g., Amazon (all three), eBay (all three).



Data Exploration Challenge

Restrictive Input Interface

- Restrictions on what queries can be issued
  - Keyword Search Interface: nothing but a set of keywords
  - Form-like Interface: only conjunctive search queries
    - e.g., List all Honda Accord cars with Price below $10,000
  - Graph Browsing Interface
    - only select one of the neighboring nodes
- We do not have complete access to the repository. No complete SQL support
  - e.g., we cannot issue "big picture" queries: e.g., SUM, MIN, MAX aggregate queries
  - e.g., we cannot issue "meta-data" queries: e.g., keyword such as DISTINCT (handy for domain discovery)



Data Exploration Challenge

Restrictive Output Interface

- Restrictions on how many tuples will be returned
  - Top-k restriction leads to three types of queries:
    - overflowing (> k): top-k elements (documents, tuples) will be selected according to a (sometimes secret) scoring function and returned
    - valid (1..k element)
    - underflowing (0 element)
  - COUNT vs. ALERT
    - An alert of overflowing can always be obtained through a web interface

> A maximum of 3000 awards are displayed. If you did not find the information you are looking for, please refine your search.

  - Page turn
    - Limited number of page turns allowed (e.g., 10-100 for Google)
      - Essentially the same as top-k restriction

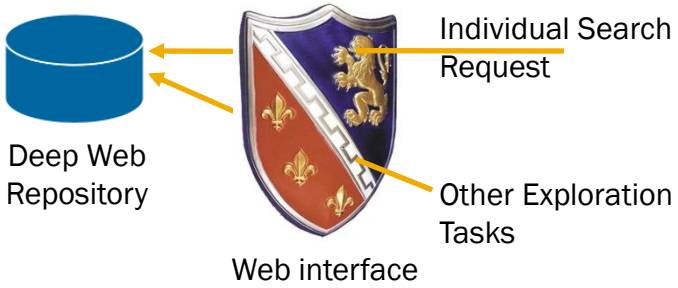
> Your search returned 41427 results. The allowed maximum number of results is 1000. Please narrow down your search criteria and try your search again.

    - Unlimited page turns
      - But a page turn also consumes a web access

1-15 of 15167 used Ford F150 cars

Data Exploration Challenge
Implications of Interface Restrictions

- Two ways to address the input/output restrictions
  - Direct negotiation with the owner of the deep web repository
    - Crawling, sampling and analytics can all be supported (if necessary)
    - Used by many real-world systems - e.g., Kayak
  - Bypass the interface restrictions
    - By issuing a carefully designed sequence of queries
    - e.g., for crawling: these queries should recall as many tuples as possible
      - or even "prove" that all tuples/documents returnable by the output interface are crawled.
    - e.g., for analytics: one should be able to infer from these queries an accurate estimation of an aggregate that cannot be directly issued because of the input interface restriction.

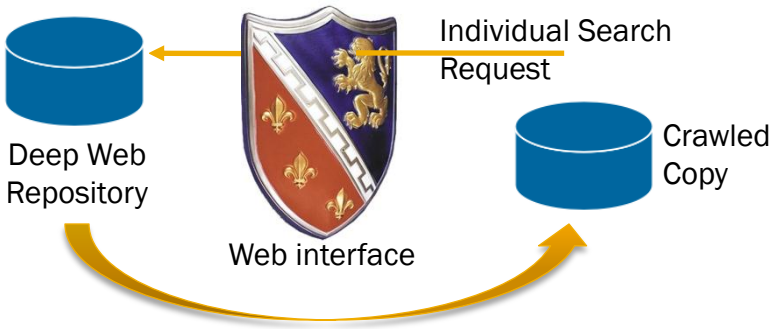


Outline

- Introduction
- Resource discovery and interface understanding
- Crawling
- Sampling

Overview of Crawling

- Motivation for crawling
  - Enable third-party web services - e.g., mash-up
  - A pre-processing step for answering queries not supported by the web interface
    - e.g., count the percentage of used cars which have GPS navigation; find all documents which contain the term "DBMS" and were last updated after Aug 1, 2011.
    - Note: these queries cannot be directly answered because of the interface restrictions.
  - Note the key differences with web crawling
- Taxonomy of crawling techniques
  - Interfaces: (a) (keyword and form-like) search interface, (b) browsing interface
  - Technical challenges: (1) find a finite set of queries that recall most if not all tuples (a challenge only for search interfaces), (2) find a small subset while maintaining a high recall, (3) issue the small subset in an efficient manner (i.e., system issues).
- Our discussion order
  - (a1), (a2), (b2), (*3)



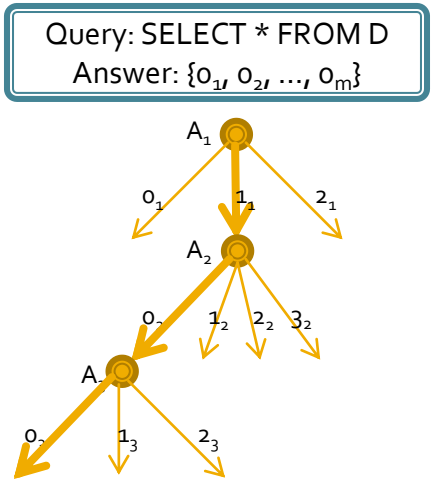
Crawling Over Search Interfaces

(a1) Find A Finite Set of Search Queries with High Recall

- Keyword search interface
  - Use a pre-determined query pool: e.g., all English words/phrases
  - Bootstrapping technique: iterative probing [CMH08]
- Form-like search interface
  - If all attributes are represented by drop-down boxes or check buttons
    - Solution is trivial 

  - If certain attributes are represented by text boxes
    - Prerequisite: attribute domain discovery 

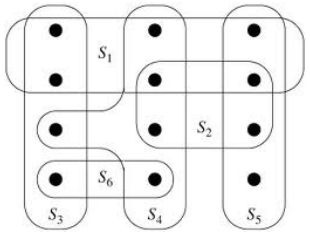
    - Nearly impossible to guarantee complete discovery [JZD11]
      - Reason: top-k restriction on output interface
      - k: $\Omega(|V|^m)$; query cost: $\Omega(m^2|V|^3)$
      - Probabilistic guarantee achievable
    - Note: domain discovery also has other applications – e.g., as a preprocessor for sampling, or standalone interest.



[CMH08] M. J. Cafarella, J. Madhavan, and A. Halevy, "Web-Scale Extraction of Structured Data", SIGMOD Record, vol. 37, 2008.
[JZD11] X. Jin, N. Zhang, G. Das, "Attribute Domain Discovery for Hidden Web Databases", SIGMOD 2011.

Crawling Over Search Interfaces
(a2) How to Efficiently Crawl

- Motivation: Cartesian product of attribute domains often orders of magnitude larger than the repository size
  - e.g., cars.com: 5 inputs, 200 million combinations vs. 650,000 tuples
- How to use the minimum number of queries to achieve a significant coverage of underlying documents/tuples
  - Essentially a set cover problem (but inputs are not properly known before hand)
- Search query selection
  - Keyword search: a heuristic of maximizing #new_elements/cost [NZC05]
    - #new_elements: not crawled by previously issued queries
    - Cost may include keyword query cost + cost for downloading details of an element
  - Form-like search: find "binding" inputs [MKK+08]
    - Informative query template: grow with increasing dimensionality
    - Good news: #informative templates grows proportionally with the database size, not #input combinations.



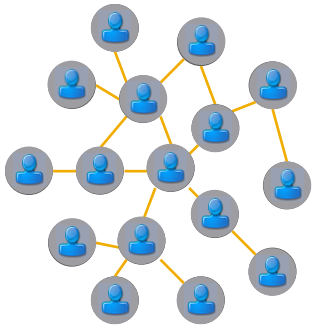
Make:Toyota
Type:Hybrid

~~Make:Jeep~~
~~Type:Hybrid~~

[NZC05] A. Ntoulas, P. Zerfos, and J. Cho, "Downloading Textual Hidden Web Content through Keyword Queries", JCDL, 2005.
[MKK+08] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy, "Google's Deep-Web Crawl", VLDB 2008.

Crawling Over Browsing Interfaces
(b2) How to Efficiently Crawl

- Technical problem
  - Hierarchical browsing: Traverse vertices of a tree
  - Graph browsing: Traverse vertices of a graph
    - Starting with a seed set of users (resp. URLs).
    - Recursively follows relationships (resp. hyperlinks) to others.
  - Exhaustive crawling vs. Focused crawling



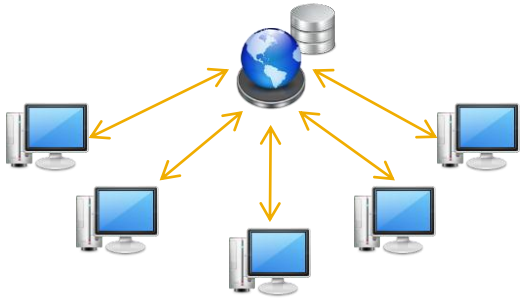
- Findings
  - Are real-world social networks indeed connected?
    - It depends – Flickr ~27%, LiveJournal ~95% [MMG+07]
  - How to select "seed(s)" for crawling?
    - Selection does not matter much as long as the number of seeds is sufficiently large (e.g., > 100) [YLW10]

[MMG+07] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks", IMC, 2007.
[YLW10] S. Ye, J. Lang, F. Wu, "Crawling Online Social Graphs", APWeb, 2010.

System Issues Related to Crawling

(*3) how to issue queries efficiently

- Using a cluster of machines for parallel crawling
  - Imperative for large-scale crawling
  - Extensively studied for web crawling
    - But are the challenges still the same for crawling deep web repositories?



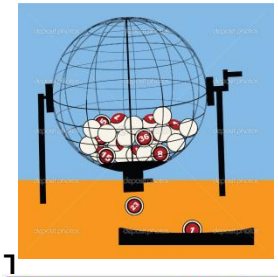
- Independent vs. Coordination
  - Overlap vs. (internal) communication overhead
  - How much coordination? Static vs. dynamic
- Politeness, or server restriction detection
  - e.g., some repositories block an IP address if queries are issued too frequently – but how to identify the maximum unblocked speed?

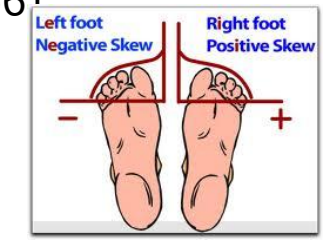
Outline

- Introduction
- Resource discovery and interface Understanding
- Crawling
- Sampling

Overview of Sampling

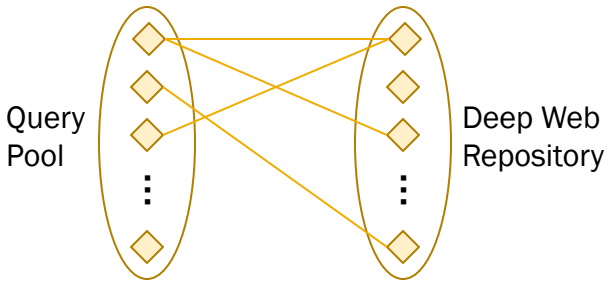
- Objective: Draw representative elements from a repository
  - Quality measure: sample skew
  - Efficiency measure: number of web accesses required
- Motivating Applications
  - Unstructured data: use sample to estimate repository sizes [SZS+06], generate content summaries [IG02], estimate average document length [BB98, BG08], etc.
    - An interesting question: Google vs. Bing, whose repository is more comprehensive?
  - Structured data: rich literature of using sampling for approximate query processing (see tutorials [Das03, GG01])
    - An interesting question: What is the average price of all 2008 Toyota Prius @ Yahoo! Autos?
  - Note (again): a sample can be later used for analytical purposes – e.g., data mining.
- Central Theme
  - Skew reduction: make the sampling distribution as close to a target distribution as possible
    - Target distribution is often the uniform distribution – in this case, the objective is to make the probability of retrieving each document as uniform as possible.





Sampling Over Keyword-Search Interfaces
Pool-Based Sampler: Basic Idea

- Query-pool based sampler
  - Assumption: there is a given (large) pool of queries which, once being issued through the web interface, can recall the vast majority of elements in the deep web repository
  - e.g., for unstructured data, a pool of English phrases
- Two types of sampling process
  - Heuristic: based on an observation that the query pool is too large to enumerate – so we have to (somehow) choose a small subset of queries (randomly or in a heuristic fashion) [IG02, SZS+06, BB98]
    - Problem: no guarantee on the "quality" (i.e., skew) of retrieved sample elements – e.g., if one randomly chooses a query and then randomly selects a document from the returned result [BB98], then longer documents will be favored over shorter ones.
  - Skew reduction: identify the source of skew and use skew-correction techniques, e.g., rejection sampling, to remove the skew.
- Interesting observation: relationship b/w keyword and sampling a bipartite graph



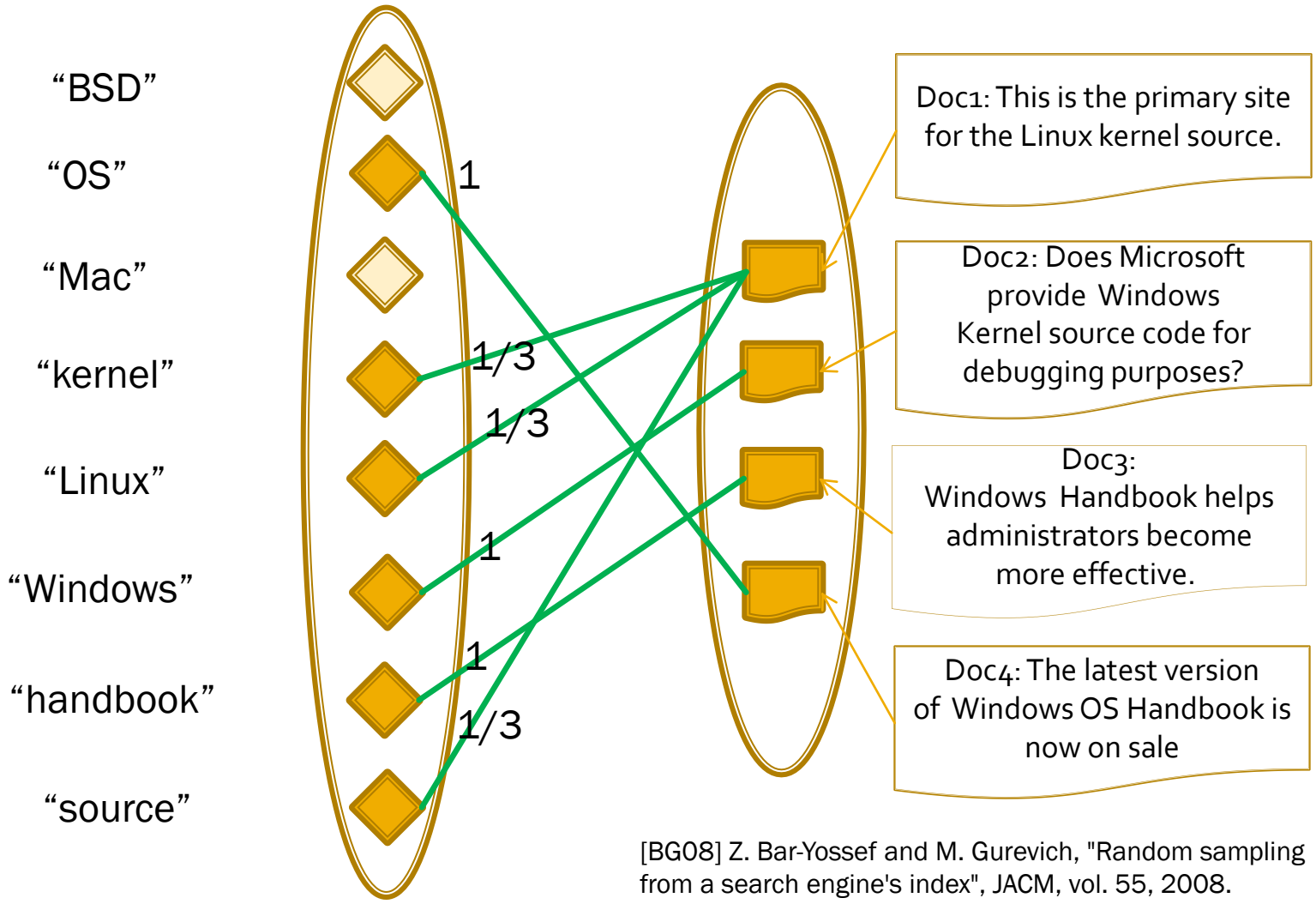
[IG02] P. G. Iperirotis and L. Gravano, "Distributed Search over the Hidden Web: Hierarchical Database Sampling and Selection", VLDB, 2002.

[SZS+06] M. Shokouhi, J. Zobel, F. Scholer, and S. Tahaghoghi, "Capturing collection size for distributed non-cooperative retrieval", SIGIR, 2006.

[BB98] K. Bharat and A. Broder, "A technique for measuring the relative size and overlap of public Web search engines", WWW, 1998.

Sampling Over Keyword-Search Interfaces

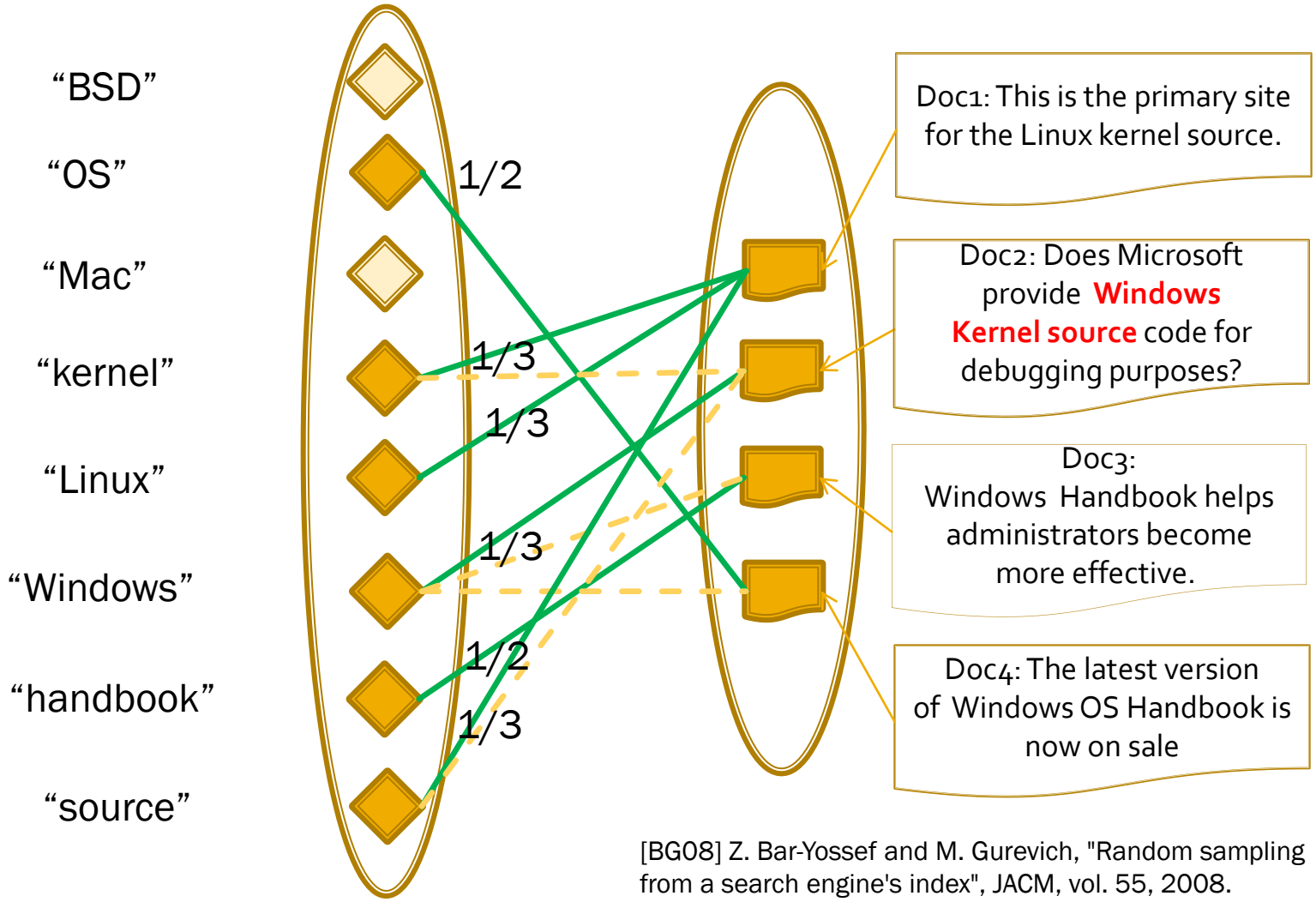
Pool-Based Sampler: Reduce Skew



[BG08] Z. Bar-Yossef and M. Gurevich, "Random sampling from a search engine's index", JACM, vol. 55, 2008.

Sampling Over Keyword-Search Interfaces

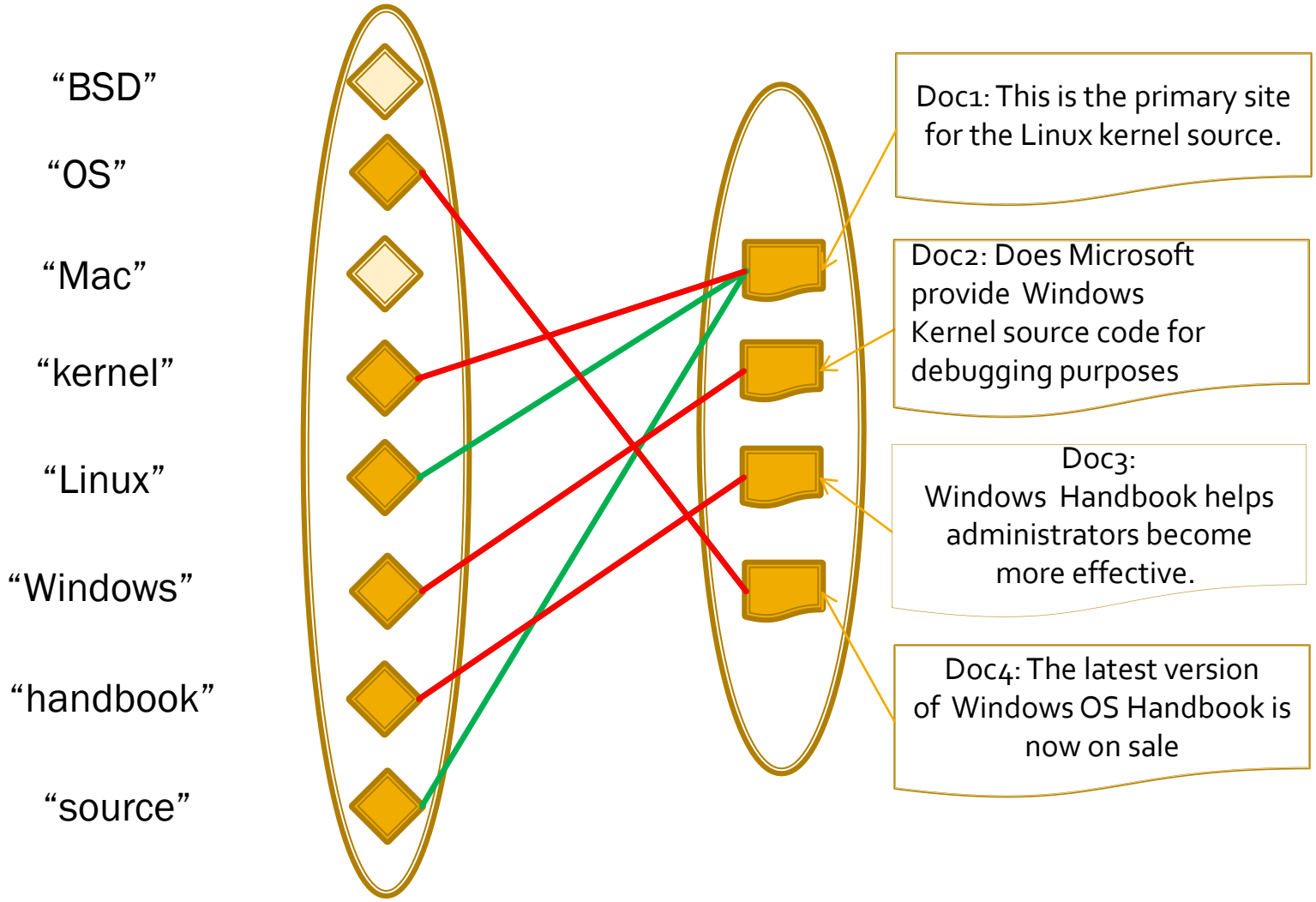
Pool-Based Sampler: Reduce Skew



[BG08] Z. Bar-Yossef and M. Gurevich, "Random sampling from a search engine's index", JACM, vol. 55, 2008.

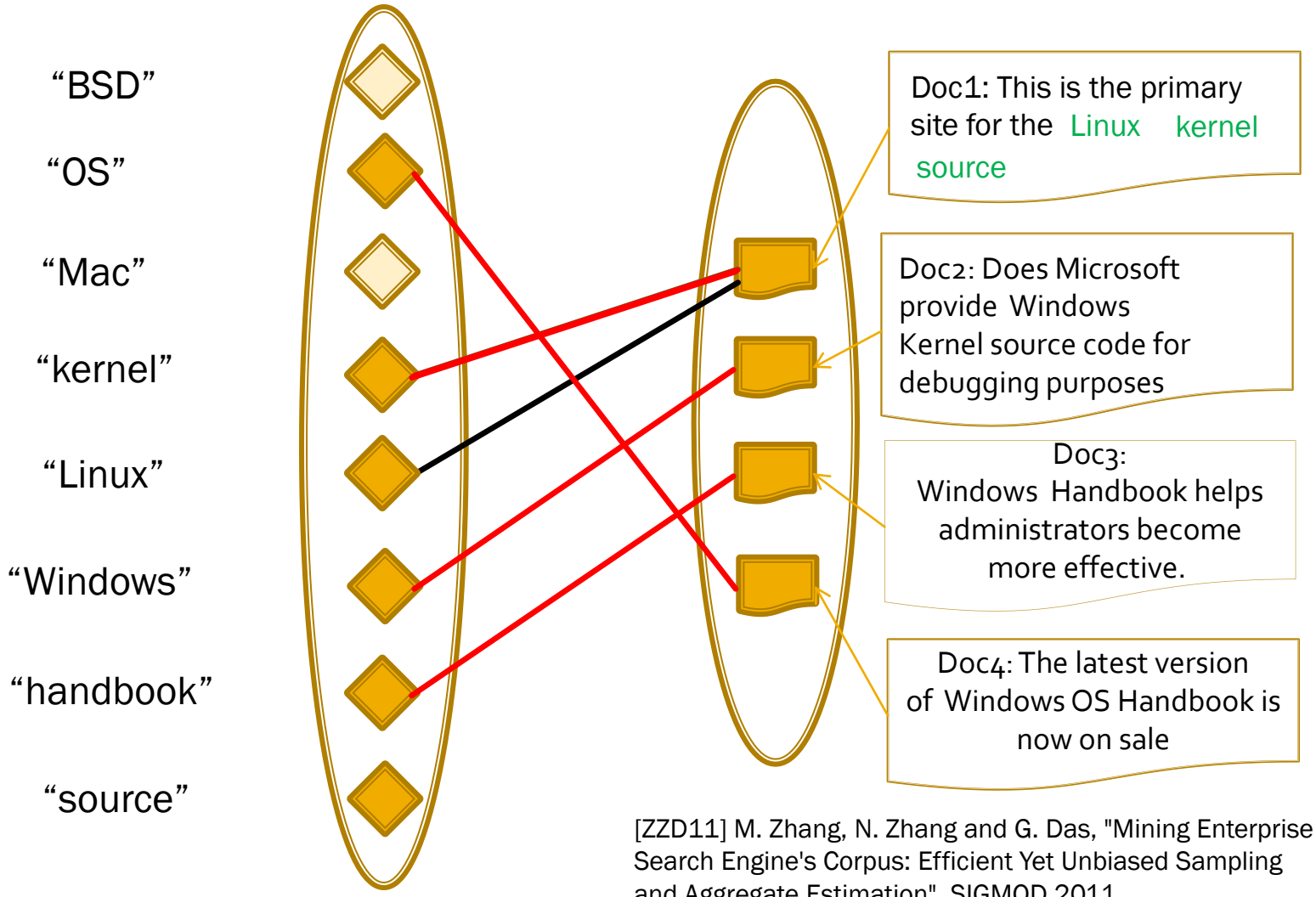
Sampling Over Keyword-Search Interfaces

Pool-Based Sampler: Remove Skew



Sampling Over Keyword-Search Interfaces

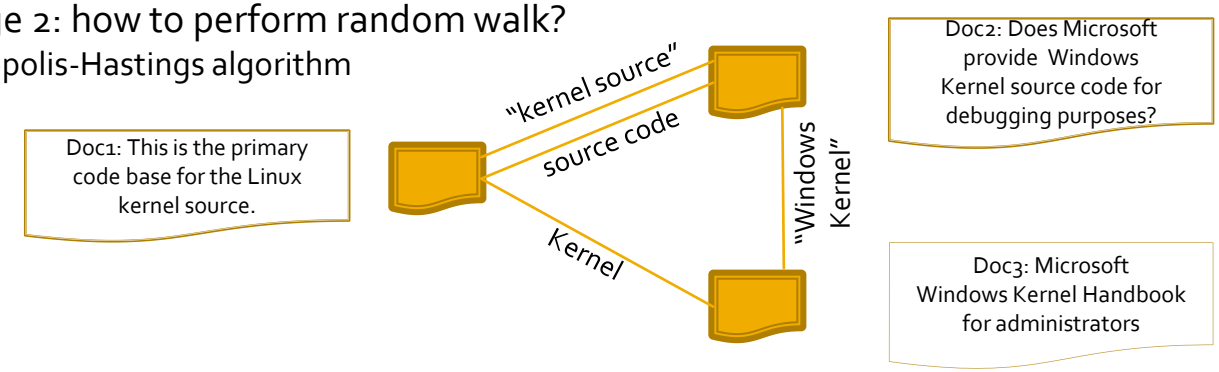
Pool-Based Sampler: Remove Skew



[ZZD11] M. Zhang, N. Zhang and G. Das, "Mining Enterprise Search Engine's Corpus: Efficient Yet Unbiased Sampling and Aggregate Estimation", SIGMOD 2011.

Sampling Over Keyword-Search Interfaces
Other Sampling Methods

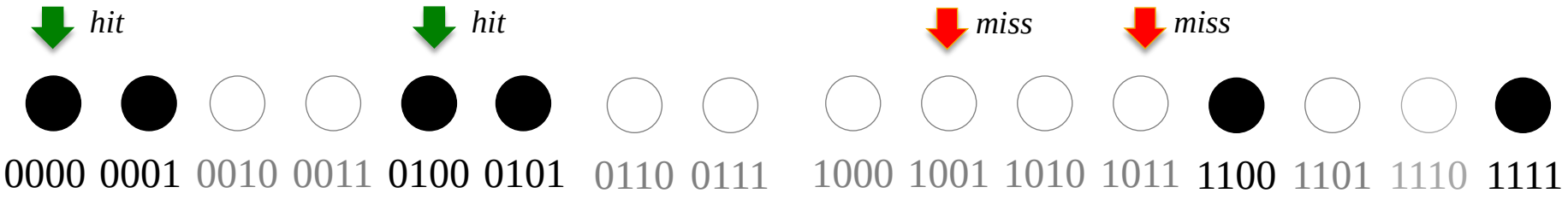
- Pool-free random walk [BG08]
  - A graph model
    - Each element in the repository is a vertex
    - Two elements are connected if they are returned by the same query
  - Random walk over the graph, two enabling factors:
    - Given an element, we can sample uniformly at random a query which returns the document. (YEA for almost all keyword search interfaces).
    - Given an element, we can find the number of queries which return the document (may incur significant query cost)
  - Challenge 1: is the graph connected?
    - Note: the set of all possible queries which might return a document can be extremely large
      - 2^n queries for a document with n words
    - Thus, we have to limit our attention to a subset of queries
      - e.g., only consecutive phrases
      - Problem: too restricted – disconnected graph, too relaxed – high cost for sampling
  - Challenge 2: how to perform random walk?
    - Metropolis-Hastings algorithm



[BG08] Z. Bar-Yossef and M. Gurevich, "Random sampling from a search engine's index", JACM, vol. 55, 2008.

Sampling Over Form-Like Interfaces
Source of Skew

- Recall: Restrictions for Form-Like Interfaces
  - Input: conjunctive search queries only
  - Output: return top-k tuples only
    (with or without the COUNT of matching tuples)
- Good News
  - Defining "designated queries" no longer a challenge
  - e.g., consider all fully specified queries – each tuple is returned by one and only one of them

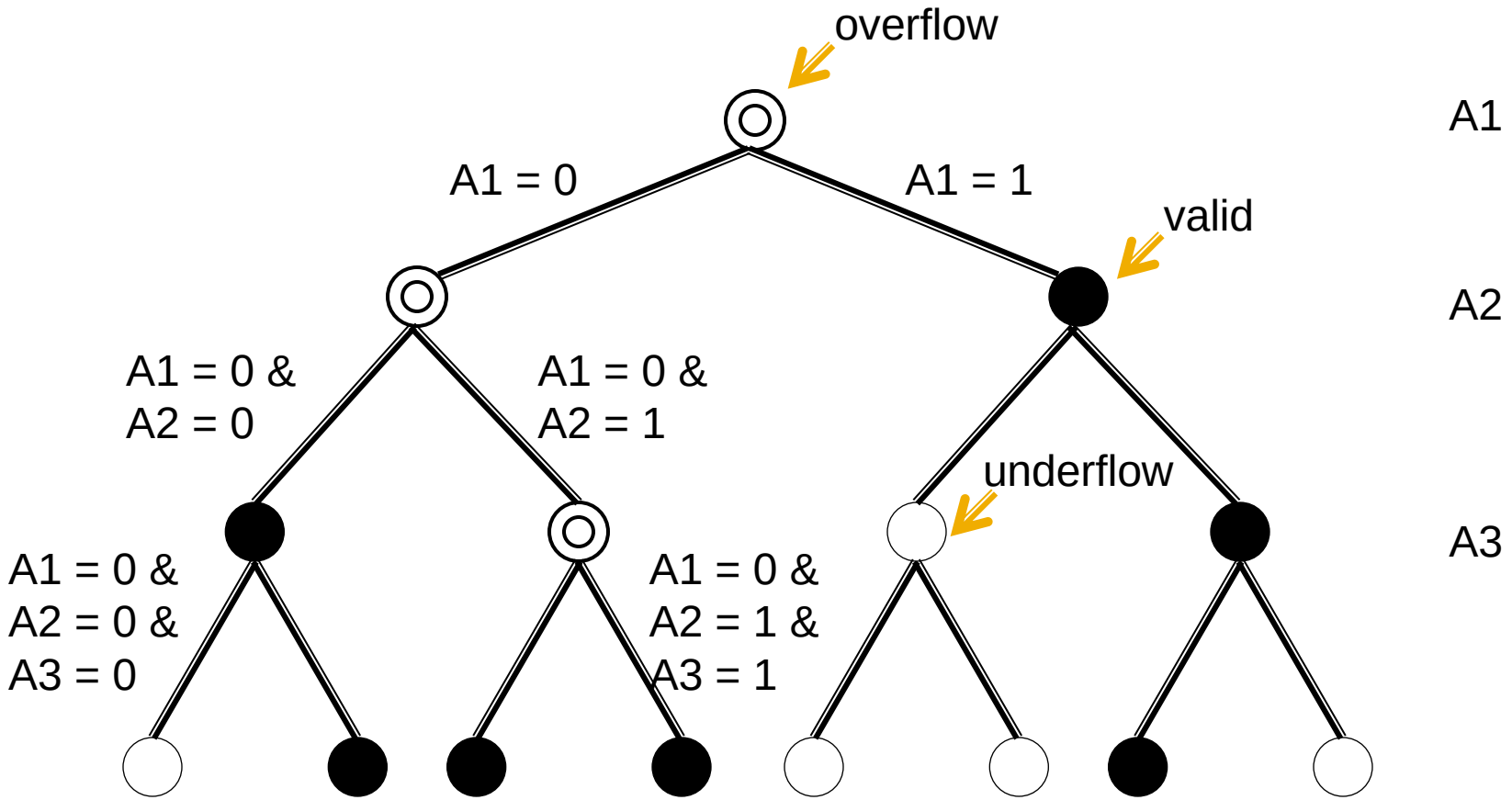


Sampling Over Form-Like Interfaces
Source of Skew

- Bad News: A New Source of Skew
  - We cannot really use fully specified queries because sampling would be really like search for a needle in a haystack
  - So we must use shorter, broader queries
    - But such queries may be affected by the top-k output restriction
    - Skew may be introduced by the scoring function used to select top-k tuples
    - e.g., skew on average price when the top-k elements are the ones with the lowest prices
- Basic idea for reducing/removing skew
  - Find non-empty queries which are not affected by the scoring function – i.e., queries which return 1 to k elements

Sampling Over Form-Like Interfaces

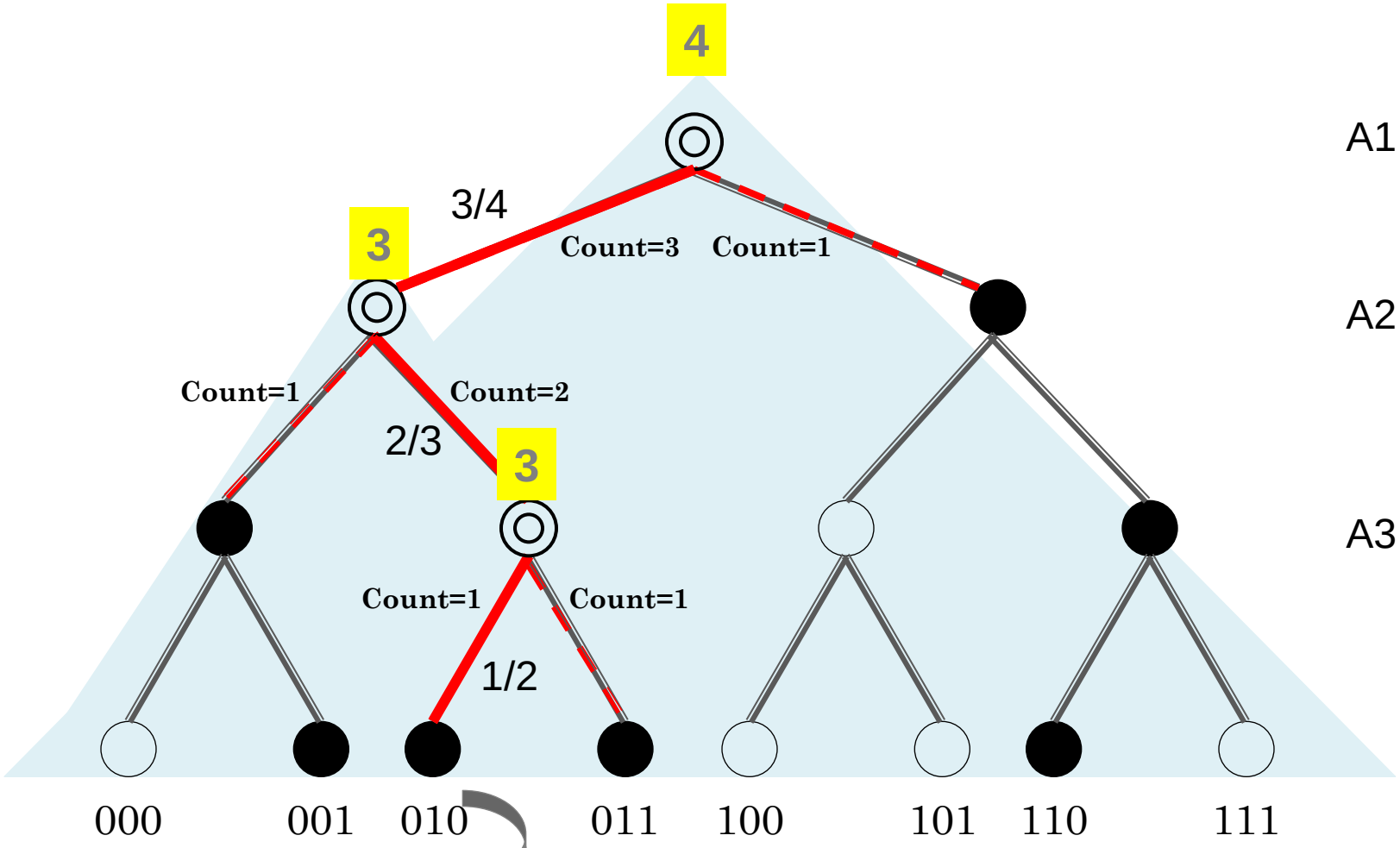
COUNT-Based Skew Removal



[DZD09] A. Dasgupta, N. Zhang, and G. Das, Leveraging COUNT Information in Sampling Hidden Databases, ICDE 2009.

Sampling Over Form-Like Interfaces

COUNT-Based Skew Removal

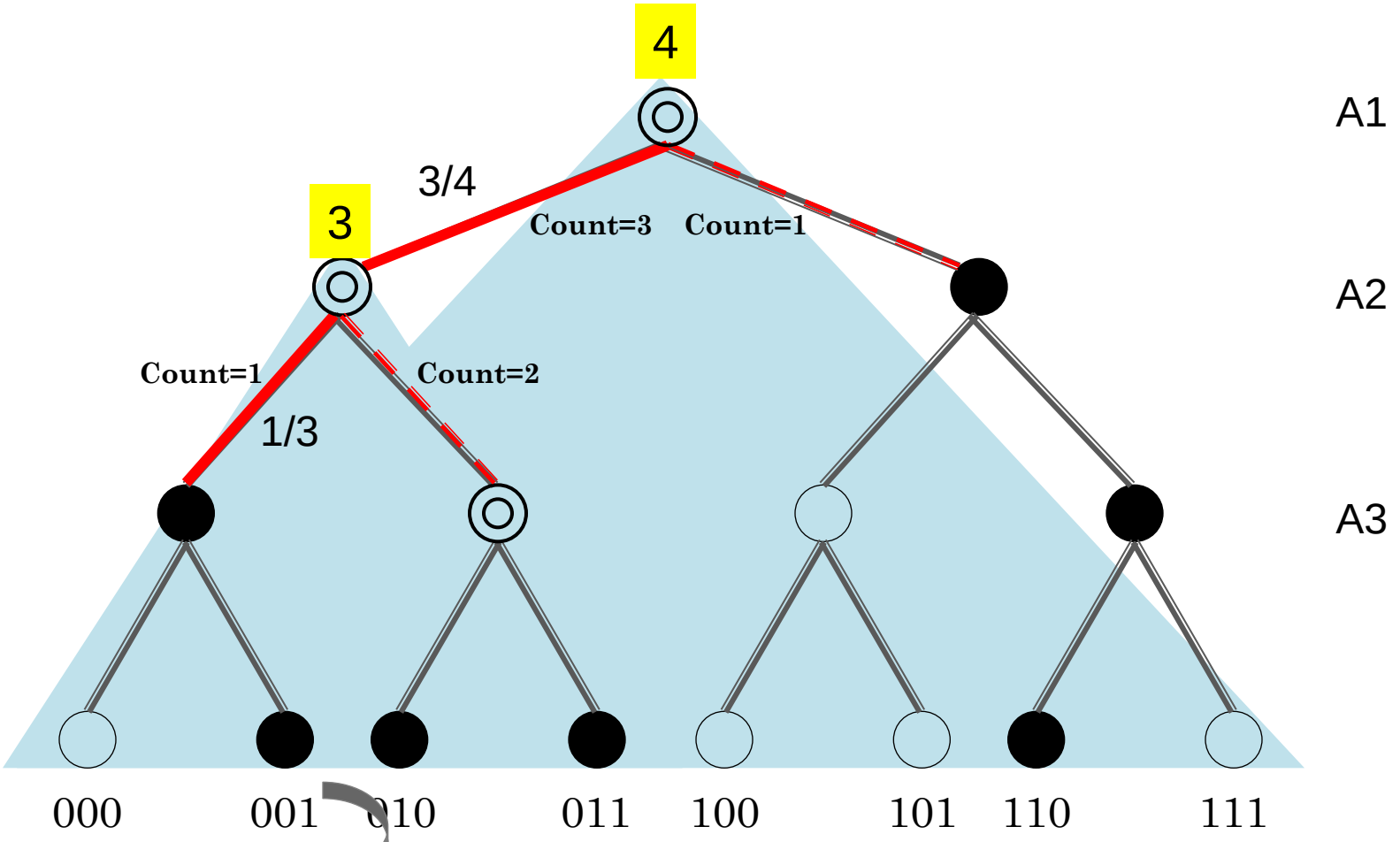


3/4 * 2/3 * 1/2 = 1/4

[DZD09] A. Dasgupta, N. Zhang, and G. Das, Leveraging COUNT Information in Sampling Hidden Databases, ICDE 2009.

Sampling Over Form-Like Interfaces

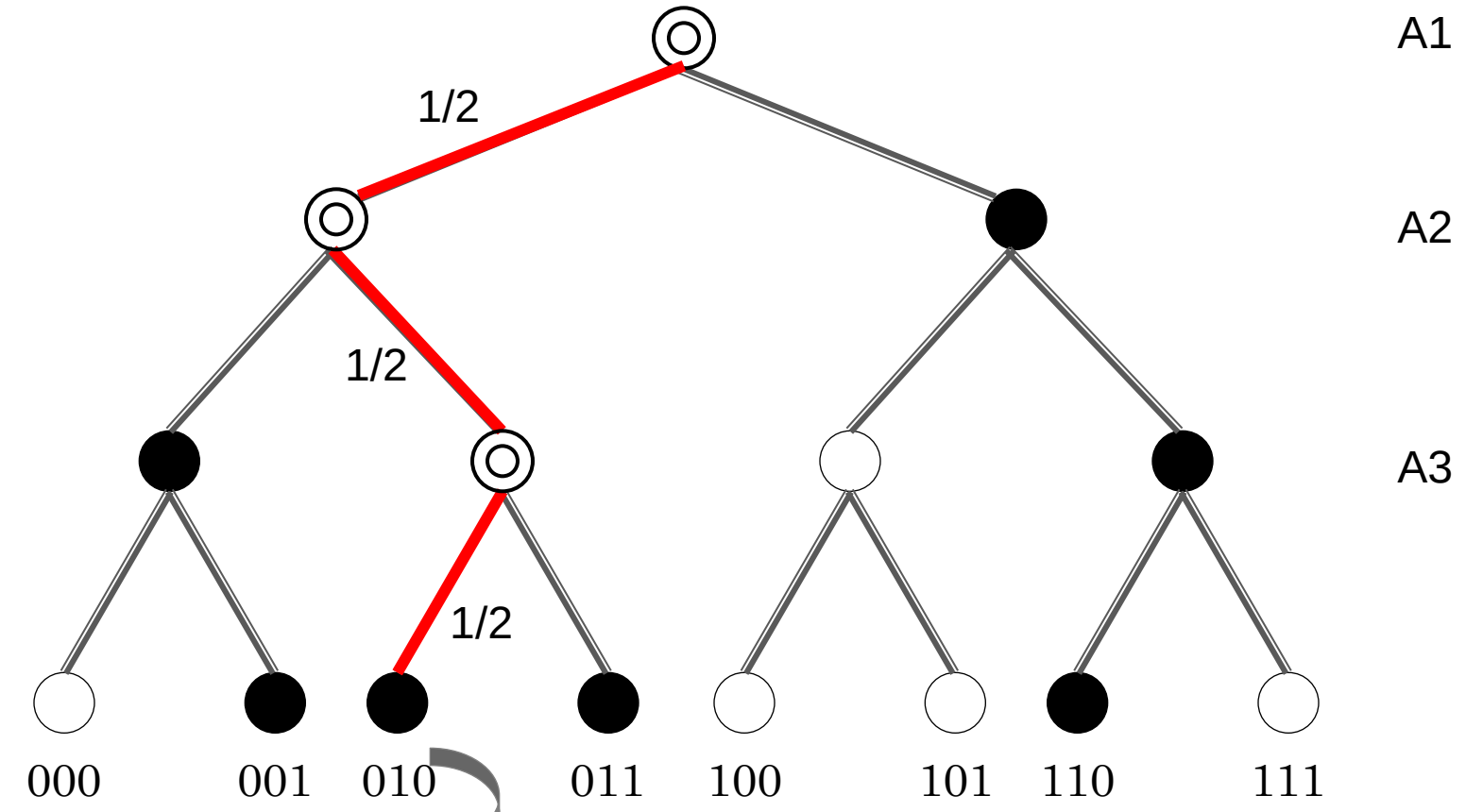
COUNT-Based Skew Removal



[DZD09] A. Dasgupta, N. Zhang, and G. Das, Leveraging COUNT Information in Sampling Hidden Databases, ICDE 2009.

Sampling Over Form-Like Interfaces

Skew Reduction for Interfaces Sans COUNT

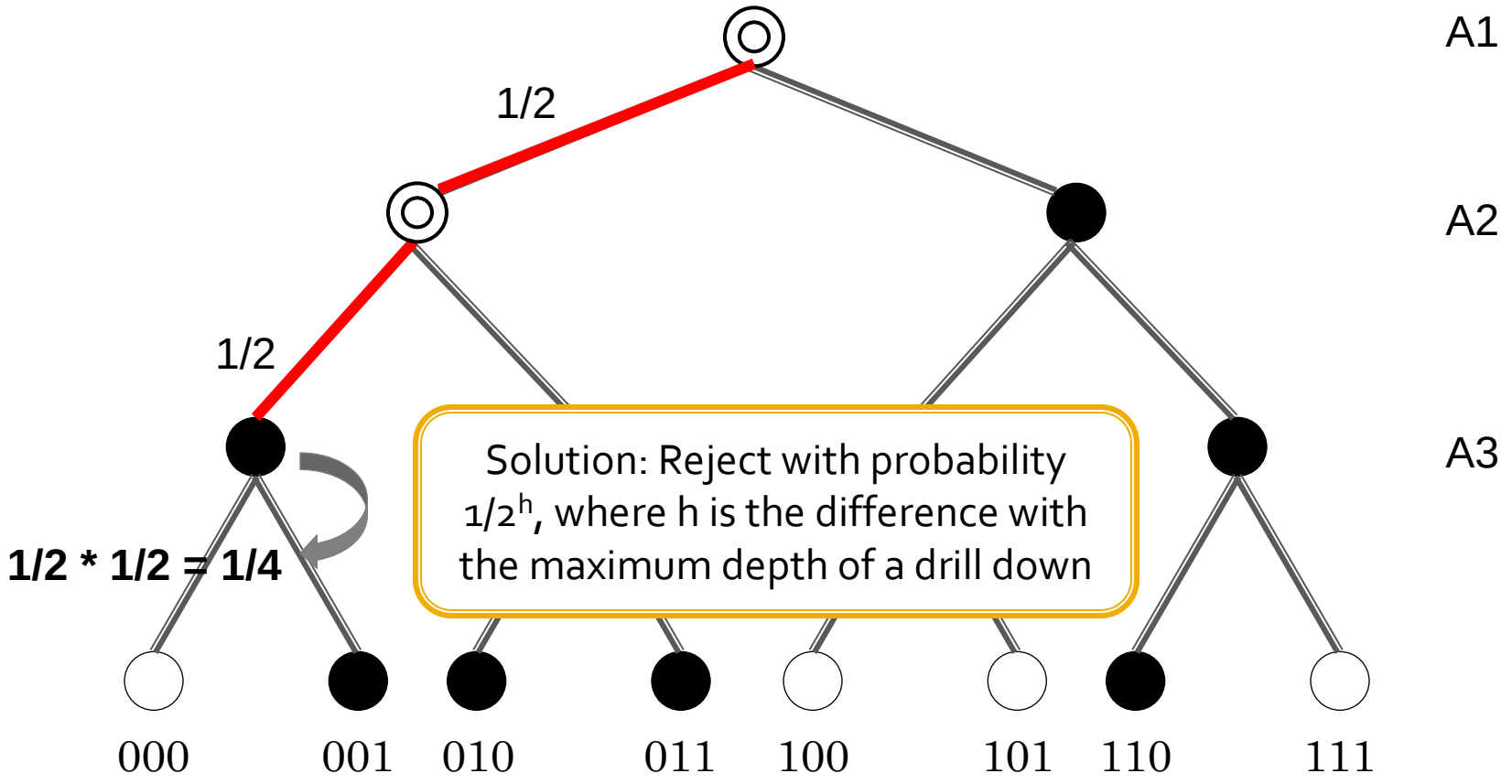


1/2 * 1/2 * 1/2 = 1/8

[DDM07] A. Dasgupta, G. Das, and H. Mannila, A Random Walk Approach to Sampling Hidden Databases, SIGMOD 2007.

Sampling Over Form-Like Interfaces

Skew Reduction for Interfaces Sans COUNT



[DDM07] A. Dasgupta, G. Das, and H. Mannila, A Random Walk Approach to Sampling Hidden Databases, SIGMOD 2007.

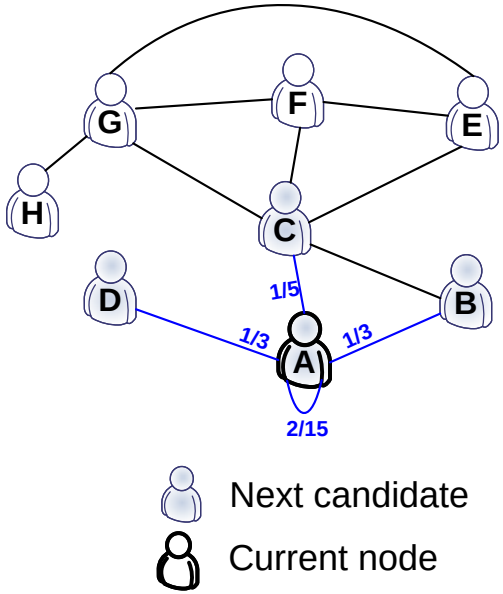
Sampling Over Graph Browsing Interfaces
Sampling by exploration

- Note: Sampling is a challenge even when the entire graph topology is given
  - Reason: Even the problem definition is tricky
    - What to sample? Vertices? Edges? Sub-graphs?
- Methods for sampling vertices, edges, or sub-graphs
  - Snowball sampling: a nonprobability sampling technique
  - Random walk with random restart
  - Forest Fire
  - …
- What are the possible goals of sampling? [LF06]
  - Criteria for a static snapshot
    - In-degree & out-degree distributions, distributions of weakly/strongly connected components (for directed graphs), distribution of singular values, clustering coefficient, etc.
  - Criteria for temporal graph evolution
    - #edges vs. #nodes over time, effective diameter of the graph over time, largest connected component size over time,

[LF06] J Leskovec and C Faloutsos, Sampling from Large Graph, KDD 2006.

Sampling Over Graph Browsing Interfaces
Unbiased Sampling

- Survey and Tutorials for random walks on graphs
  - [Lov93], [LF08], [Mag08]
- Simple random walk is inherently biased
  - Stationary distribution: each node v has probability of d(v)/(2|E|) of being selected, where d(v) is the degree of v and |E| is the total number of edges – i.e., p(v) ~ d(v)
- Skew correction
  - Re-weighted random walk [VH08]
    - Rejection sampling
    - Or, if the objective is to use the samples to estimate an aggregate, then apply Hansen-Hurwitz estimator after a simple random walk.
  - Metropolis-Hastings random walk [MRR+53]
    - Transition probability from u to its neighbor v: min(1, d(u)/d(v))/d(u)
    - Stay at u with the remaining probability
    - Leading to a uniform stationary distribution



Example taken from the slides of M Gjoka, M Kurant, C Butts, A Markopoulou, "Walking in Facebook: Case Study of Unbiased Sampling of OSNs", INFOCOM 2010

[Mag08] M. Maggioni, Tutorial - Random Walks on Graphs Large-time Behavior and Applications to Analysis of Large Data Sets, MRA 2008.
[LF08] J. Leskovec and C. Faloutsos, "Tools for large graph mining: structure and diffusion", WWW (Tutorial), 2008.
[Lov93] L. Lovasz, "Random walks on graphs: a survey", Combinatorics, Paul Erdos is Eighty, 1993.
[VH08] E. Volz and D. Heckathorn, "Probability based estimation theory for respondent-driven sampling," J. Official Stat., 2008.
[MRR+53] N. Metropolis, M. Rosenblut, A. Rosenbluth, A. Teller, and E. Teller, Equation of state calculation by fast computing machines, J. Chem. Phys., vol. 21, 1953.

Outline

- Introduction
- Resource discovery and interface understanding
- Crawling
- Sampling

for **data analytics** and the uncut version see the whole tutorial