



Unbiased Estimation of Size and Other Aggregates Over Hidden Databases



Arjun Dasgupta

University of Texas at Arlington

Xin Jin

George Washington University

Bradley Jewell

University of Texas at Arlington

Nan Zhang

George Washington University

Gautam Das

University of Texas at Arlington

Hidden Databases on the Web



Deep Web includes dynamic contents, unlinked pages, private web, contextual web, which cannot be extracted by standard crawlers. Its size is estimated to be 400 to 550 times the size of commonly defined WWW.

Make: Ford Year: 2000 To: 2006 Mileage: Any From: Your City ZIP: 76102 Listing Type: Used

Model: F150 Price: \$1,000 To: \$10,000 Distance: 3 miles To: 2002 For Sale By: All Sellers



Title	Make	Model	Year	Condition	Mileage	Location	Price
2003 Ford F150	Ford	F150	2003	Used	86,205	Alexandria, Virginia, 22307	\$9,990
2003 Ford F150 XLT	Ford	F150	2001	Used	95,645	102 National Dr, Fredericksburg, VA, 22406, United States	\$4,900
Ford F150	Ford	F150	2002	Used	58,034	Carport Network Dealer, Arroyo, MD, 21401, USA	\$7,997
2001 Ford F150 XLT	Ford	F150	2001	Used	62,116	7327 Ritten Highway, Clarysville, MD, 21041, USA	\$7,999
2001 Ford F150 XLT	Ford	F150	2006	Used	49,900	402 Centex Drive, Mount Airy, NC, 27571, USA	\$9,999
2001 Ford F150	Ford	F150	2001	Used	102,810	14548 Jefferson Davis Highway Woodbridge, VA, 22191	\$4,999
2001 Ford F150	Ford	F150	2004	Used	66,433	3716 Deer Highway, Upper Marlboro, MD, 20772	\$7,999
2001 Ford F150 XLT	Ford	F150	2004	Used	101,644	Alexandria, MD 21401	\$7,999
2001 Ford F150 Superduty L1	Ford	F150	2001	Used	112,432	1029 Marwood Avenue, Alexandria, VA, 22302	\$9,990
2001 Ford F150	Ford	F150	2004	Used	66,433	Upper Marlboro, Maryland, 20772	\$9,999
2001 Ford F150	Ford	F150	2001	Used	79,643	Leesburg, VA 20176	\$9,990
2001 Ford F150	Ford	F150	2001	Used	117,470	Washington, DC 20011	\$9,990

Hidden Databases
 + Large part of deep web
 + Dynamic content generated by proprietary data
 + Top-k constraint based data delivery model serving user queries

😊 YES
 🙄 NO

👉 Tuple level search queries
 🚫 Aggregate queries

Estimation of Aggregates over Hidden Databases

Why Aggregates Matter to Clients/External users

- + Data quality, freshness, content bias, size, etc.
- + Verify size/quality claims by providers

Examples of Aggregate Queries on a Hidden Database

- + COUNT total # of "passenger" cars in a used car database
- + AVG(Mileage), SUM(Price) of "Honda" cars in "Dallas"

Our Objective

+ **ACCURATELY** and **EFFICIENTLY** estimate size and other aggregates of a hidden database (w/ or w/o selection conditions) by issuing a small number of queries through its public interface

Challenges

- + Top-k constraint
- + Existing samplers cannot be easily adapted

Baseline Techniques

Brute Force Method: $\tilde{m} = \frac{h}{H} \times Dom$; $E(\tilde{m}) = E(\frac{h}{H}) \times Dom = m$;

where h is the number of tuples returned from H fully specified random queries, \tilde{m} is the estimated database size of the true size m .

+ Need **at least** $\frac{Dom}{m}$ queries to be issued

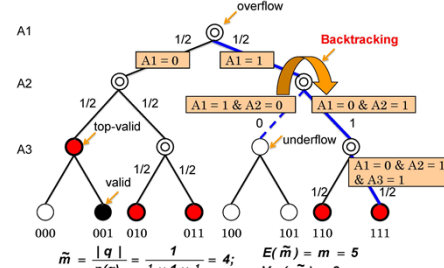
Capture/Recapture Method: samples a population and estimates the size according to the recaptured counts of entities being sampled.

$$\tilde{m} = \frac{|C1| \times |C2|}{|C1 \cap C2|}$$

where $|C1|$ is the number of tuples in the first sample and $|C1 \cap C2|$ is the number of tuples appearing in both samples.

- + Introduce bias
- + Need **at least** \sqrt{m} queries to be issued

Unbiased Estimation via Random Walk with Backtracking



Why Our Estimator is Unbiased

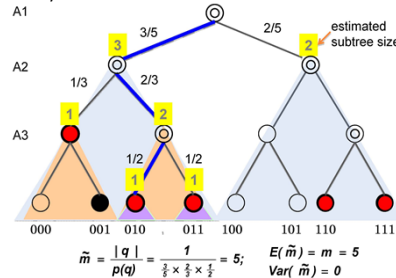
$$E(\tilde{m}) = \sum_{q \in \Omega_{TV}} \frac{|q|}{p(q)} \times p(q) = \sum_{q \in \Omega_{TV}} |q| = m;$$

Ω_{TV} : the set of all top-valid nodes in the query tree.

Backtracking: ensures that every random walk returns an estimation.

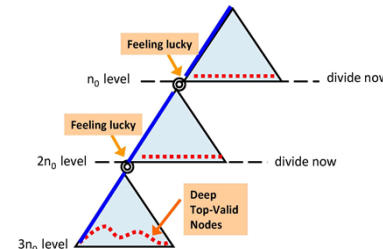
Techniques for Variance Reduction

Weight Adjustment: adjusting alignment between branch selection probability and the distribution of the measure attribute.

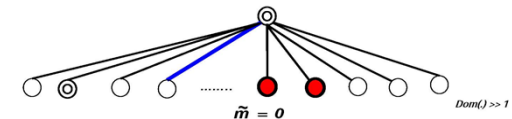


A major source of variance results from deep large subtrees.

Divide & Conquer: ensures that as long as we get lucky once, such a subtree can be thoroughly explored



For Categorical Data



- + When terminating at an underflowing node, the random walk returns a zero-sized estimation (NO backtracking).
- + Variance reduction: both weight adjustment and divide & conquer are still valid.

Generic Aggregates Extension

- + Extend to aggregate queries like SUM (unbiased) and AVERAGE (biased for this case)

+ Generic aggregates extension: replaces $\frac{|q|}{p(q)}$ by $\frac{AGG(q)}{p(q)}$

Experimental Results

Datasets

- + Boolean synthetic: 200,000 tuples, 40 boolean attributes, iid vs. mix
- + Offline Yahoo! Auto: 15, 211 tuples, 32 boolean and 6 categorical attributes
- + Online Yahoo Auto: http://autos.yahoo.com/listings/advanced_search

Performance Metrics

- + Query cost: the number of queries issued through front-end interfaces
- + MSE: mean square error
- + relative error, i.e. $\frac{\tilde{m}}{m}$ where $\frac{|\tilde{m} - m|}{\tilde{m}}$ is the estimated value of the original m

