

Google™



Hyperlink Analysis on the Web

Monika Henzinger  
[monika@google.com](mailto:monika@google.com)

# Outline

---

- Random Walks
- Classic Information Retrieval (IR) vs Web IR
- Hyperlink Analysis:
  - PageRank
  - HITS
- Random Walks on the Web

# Random Walks

---

- **Random Walk** = discrete-time stochastic process over a graph  $G=(V,E)$  with a transition probability matrix  $P$ 
  - Random Walk is at one node at any time, making node-transitions at time steps  $t=1,2, \dots$  with  $P_{ij}$  being the probability of going to node  $j$  when at node  $i$
  - Initial node chosen according to some probability distribution  $q^{(0)}$  over  $S$

# Random Walks (cont.)

---

- $q^{(t)}$  = row vector whose  $i$ -th component is the probability that the chain is in node  $i$  at time  $t$
- $q^{(t+1)} = q^{(t)} P \Rightarrow q^{(t)} = q^{(0)} P^t$
- A **stationary distribution** is a probability distribution  $q$  such that  $q = q P$  (steady-state behavior)
- Example:
  - $P_{ij} = 1/\text{degree}(i)$  if  $(i,j)$  in  $G$  and 0 otherwise, then  $q_i = \text{degree}(i)/2m$

# Random Walks (cont.)

---

- Theorem: Under certain conditions:
  - There exists a unique stationary distribution  $q$  with  $q_i > 0$  for all  $i$
  - Let  $N(i,t)$  be the number of times the random walk visits node  $i$  in  $t$  steps. Then, the fraction of steps the walk spends at  $i$  equals  $q_i$ , i.e.

$$\lim_{t \rightarrow \infty} \frac{N(i,t)}{t} = q_i$$

# Information Retrieval

---

- **Input:** Document collection
- **Goal:** Retrieve documents or text with information **content** that is **relevant** to user's **information need**
- **Two aspects:**
  1. Processing the collection
  2. Processing queries (searching)

# Classic information retrieval

---

- Ranking is a function of *query term frequency within the document (tf)* and *across all documents (idf)*
- This works because of the following **assumptions** in classical IR:
  - Queries are **long and well specified**  
“What is the impact of the Falklands war on Anglo-Argentinean relations”
  - Documents (e.g., newspaper articles) are **coherent, well authored**, and are usually about one topic
  - The **vocabulary is small** and relatively well understood

# Web information retrieval

---

- **None of these assumptions hold:**
  - Queries are **short**: 2.35 terms in avg
  - **Huge variety in documents**: language, quality, duplication
  - Huge **vocabulary**: 100s million of terms
  - **Deliberate misinformation**
- Ranking is a function of the *query terms and of the hyperlink structure*

# Hyperlink analysis

---

- Idea: Mine structure of the *web graph*
  - Each web page is a node
  - Each hyperlink is a directed edge
- Related work:
  - Classic IR work (citations = links) a.k.a. “Bibliometrics” [K’63, G’72, S’73,...]
  - Socio-metrics [K’53, MMSM’86,...]
  - Many Web related papers use this approach [PPR’96, AMM’97, S’97, CK’97, K’98, BP’98,...]

# Google's approach

---

- Assumption: A **link** from page A to page B is a **recommendation** of page B by the author of A (we say B is *successor* of A)
    - ➔ Quality of a page is related to its in-degree
  - Recursion: Quality of a page is related to
    - its in-degree, and to
    - the *quality* of pages linking to it
- ➔ **PageRank** [BP '98]

# Definition of PageRank

---

- Consider the following infinite **random walk** (surf):
  - Initially the surfer is at a random page
  - At each step, the surfer proceeds
    - to a randomly chosen web page with probability  $d$
    - to a randomly chosen successor of the current page with probability  $1-d$
- **The PageRank of a page  $p$  is *the fraction of steps the surfer spends at  $p$  in the limit.***

# PageRank (cont.)

---

By previous theorem:

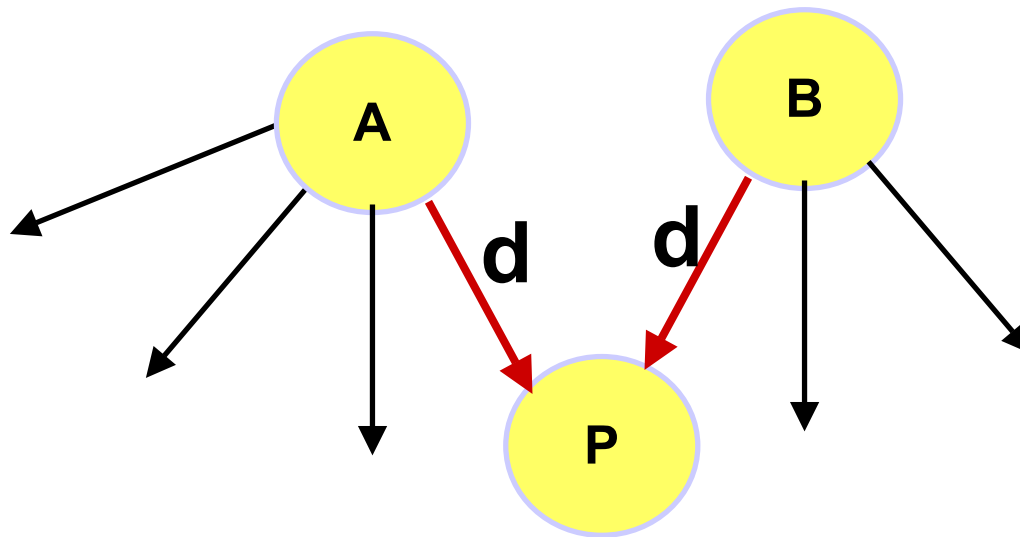
- PageRank = stationary probability for this Markov chain, i.e.

$$PageRank(p) = \frac{d}{n} + (1-d) \sum_{(q,p) \in E} PageRank(q) / outdegree(q)$$

where  $n$  is the total number of nodes in the graph

# PageRank (cont.)

---



PageRank of P is

$$(1-d) * \left( \frac{1}{4} \text{th the PageRank of A} + \frac{1}{3} \text{rd the PageRank of B} \right) + d/n$$

# PageRank

---

- Used in Google's ranking function
- Query-independent
- Summarizes the “web opinion” of the page importance

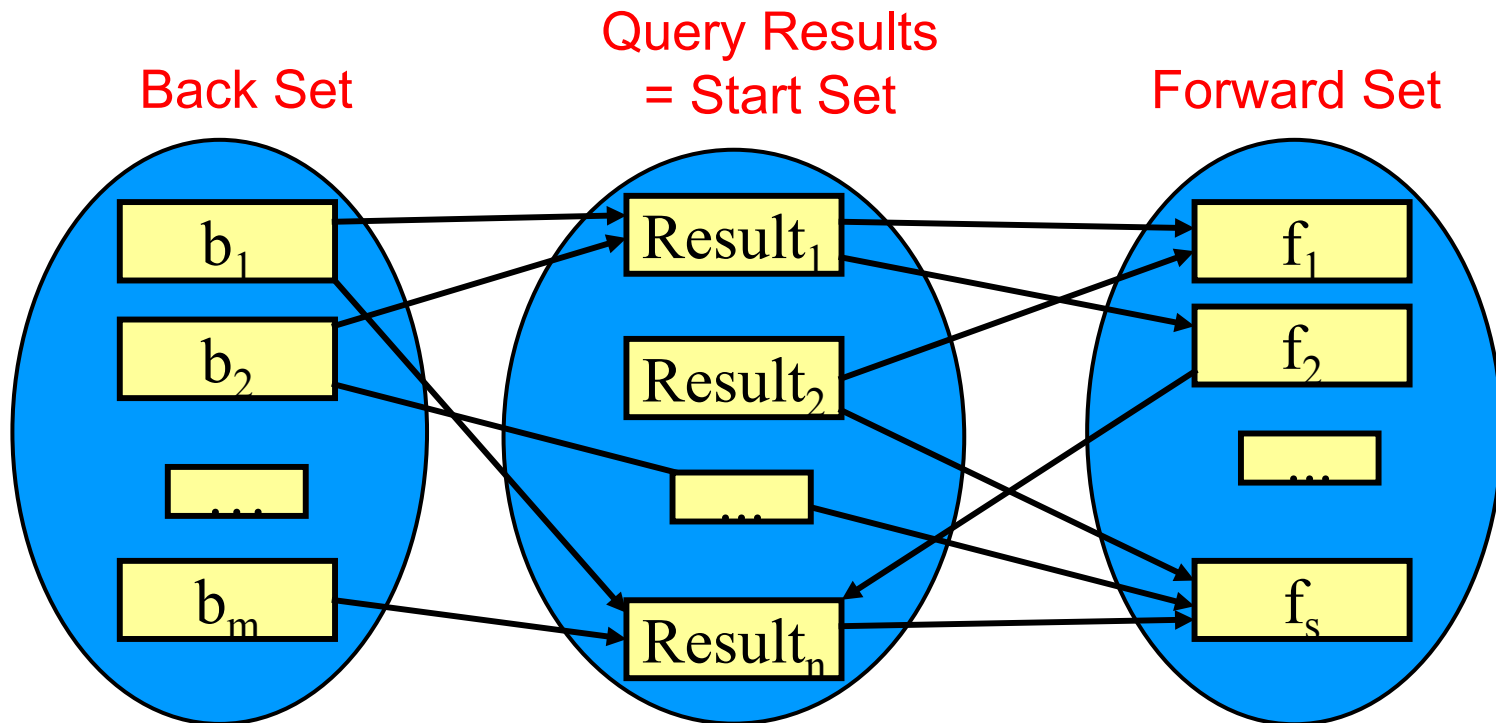
# Outline

---

- Markov Chains and Random Walks
- Information Retrieval (IR) vs Web IR
- Hyperlink Analysis:
  - PageRank
  - HITS
- Random Walks on the Web

# Neighborhood graph

- Subgraph associated to each query



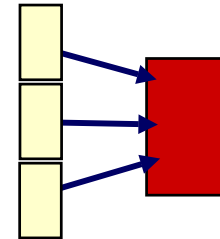
An edge for each hyperlink, but no edges within the same host

# HITS [K'98]

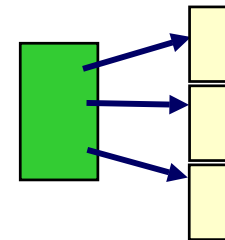
---

- **Goal:** Given a query find:

- Good sources of content (authorities)



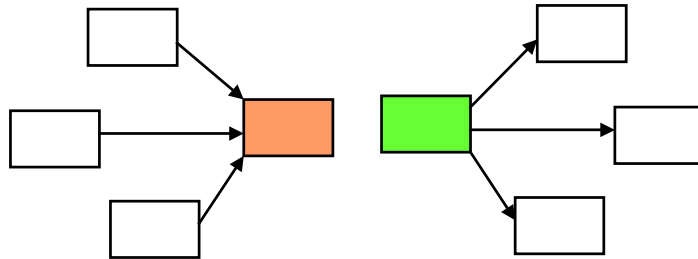
- Good sources of links (hubs)



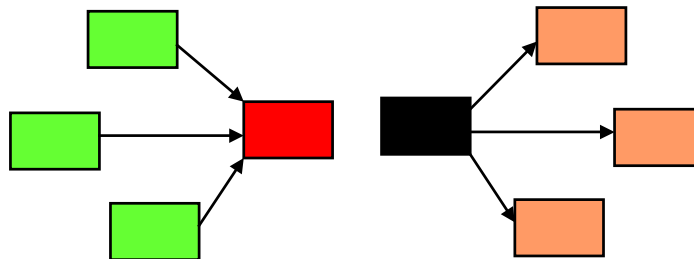
# Intuition

---

- **Authority** comes from in-edges.  
Being a **good hub** comes from out-edges.



- **Better authority** comes from in-edges from **good hubs**. Being a **better hub** comes from out-edges to **good authorities**.



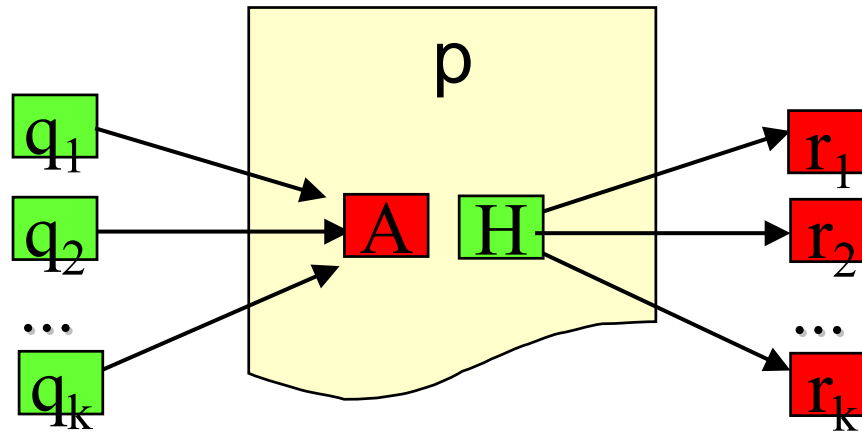
# HITS details

Repeat until  $\overrightarrow{HUB}$  and  $\overrightarrow{AUTH}$  converge:

Normalize  $\overrightarrow{HUB}$  and  $\overrightarrow{AUTH}$

$HUB[p] := \sum AUTH[r_i]$  for all  $r_i$  with  $(p, r_i)$  in  $E$

$AUTH[p] := \sum HUB[q_i]$  for all  $q_i$  with  $(q_i, p)$  in  $E$



# PageRank vs. HITS

---

- Computation:
    - Once for all documents and queries (offline)
  - Query-independent – requires combination with query-dependent criteria
  - Hard to spam
- Computation:
    - Requires computation for each query
  - Query-dependent
  - Relatively easy to spam
  - Quality depends on quality of start set
  - Gives hubs as well as authorities

# PageRank vs. HITS

---

- [Lempel] Not rank-stable:  $O(1)$  changes in graph can change  $O(N^2)$  order-relations
- [Ng,Zheng, Jordan01] “Value”-Stable: change in  $k$  nodes (with PR values  $p_1, \dots, p_k$ ) results in  $p^*$  s.t.

$$\| p^* - p \| \leq 2 \sum_{j=1}^k p_j / d$$

- Not rank-stable
- “value”-stability depends on gap  $g$  between largest and second largest eigenvector: change of  $O(g)$  nodes results in  $p^*$  s.t.

$$\| p^* - p \| = \Omega(1)$$

# Outline

---

- Random Walks
- Classic Information Retrieval (IR) vs Web IR
- Hyperlink Analysis:
  - PageRank
  - HITS
- Random Walks on the Web

# Let's do it!

---

- Perform PageRank random walk
- Select uniform random sample from resulting pages

- “Quality-biased” sample of the web
- Useful for estimation:
  - Web properties: Percentage of *high-quality* pages in a domain, in a language, on a topic, ...
  - Search engine comparison: Sum of probabilities of pages in the index (**index quality**)

# Sampling pages (almost) according to PageRank

---

- **Problems:**
  - Starting state bias: finite walk only approximates PageRank.
  - Can't jump to a random page; instead, jump to a random page on a random host seen so far.
- Sampling pages according to a distribution that behaves similarly to PageRank

# Experiments on the real web

---

- Performed two long random walks with  $d=1/7$  starting at [www.yahoo.com](http://www.yahoo.com)

	<b>Walk 1</b>	<b>Walk2</b>
length	18 hours	54 hours
HTML pages successfully downloaded	1,393,265	2,940,794
unique HTML pages	509,279	1,002,745
sampled pages	1,025	1,100

# Random walk effectiveness

---

- **Repeatability:** Index quality results are consistent over the 2 walks
- **Reduction of initial bias:** Bias for [www.yahoo.com](http://www.yahoo.com) is reduced in longer walk
- **Similarity to PageRank:**
  - Pages (or hosts) that are “highly-reachable” are visited often by the random walks
  - The average indegree of pages with indegree  $\leq 1000$  is high:
    - 53 in walk 1
    - 60 in walk 2

# Most frequently visited pages

---

Page	Freq. Walk2	Freq. Walk1	Rank Walk1
www.microsoft.com/	3172	1600	1
www.microsoft.com/windows/ie/default.htm	2064	1045	3
www.netscape.com/	1991	876	6
www.microsoft.com/ie/	1982	1017	4
www.microsoft.com/windows/ie/download/	1915	943	5
www.microsoft.com/windows/ie/download/all.htm	1696	830	7
www.adobe.com/prodindex/acrobat/readstep.htm	1634	780	8
home.netscape.com/	1581	695	10
www.linkexchange.com/	1574	763	9
www.yahoo.com/	1527	1132	2

# Most frequently visited hosts

---

Site	Frequency Walk 2	Frequency Walk 1	Rank Walk 1
www.microsoft.com	32452	16917	1
home.netscape.com	23329	11084	2
www.adobe.com	10884	5539	3
www.amazon.com	10146	5182	4
www.netscape.com	4862	2307	10
excite.netscape.com	4714	2372	9
www.real.com	4494	2777	5
www.lycos.com	4448	2645	6
www.zdnet.com	4038	2562	8
www.linkexchange.com	3738	1940	12
www.yahoo.com	3461	2595	7

# Estimating search engine index quality

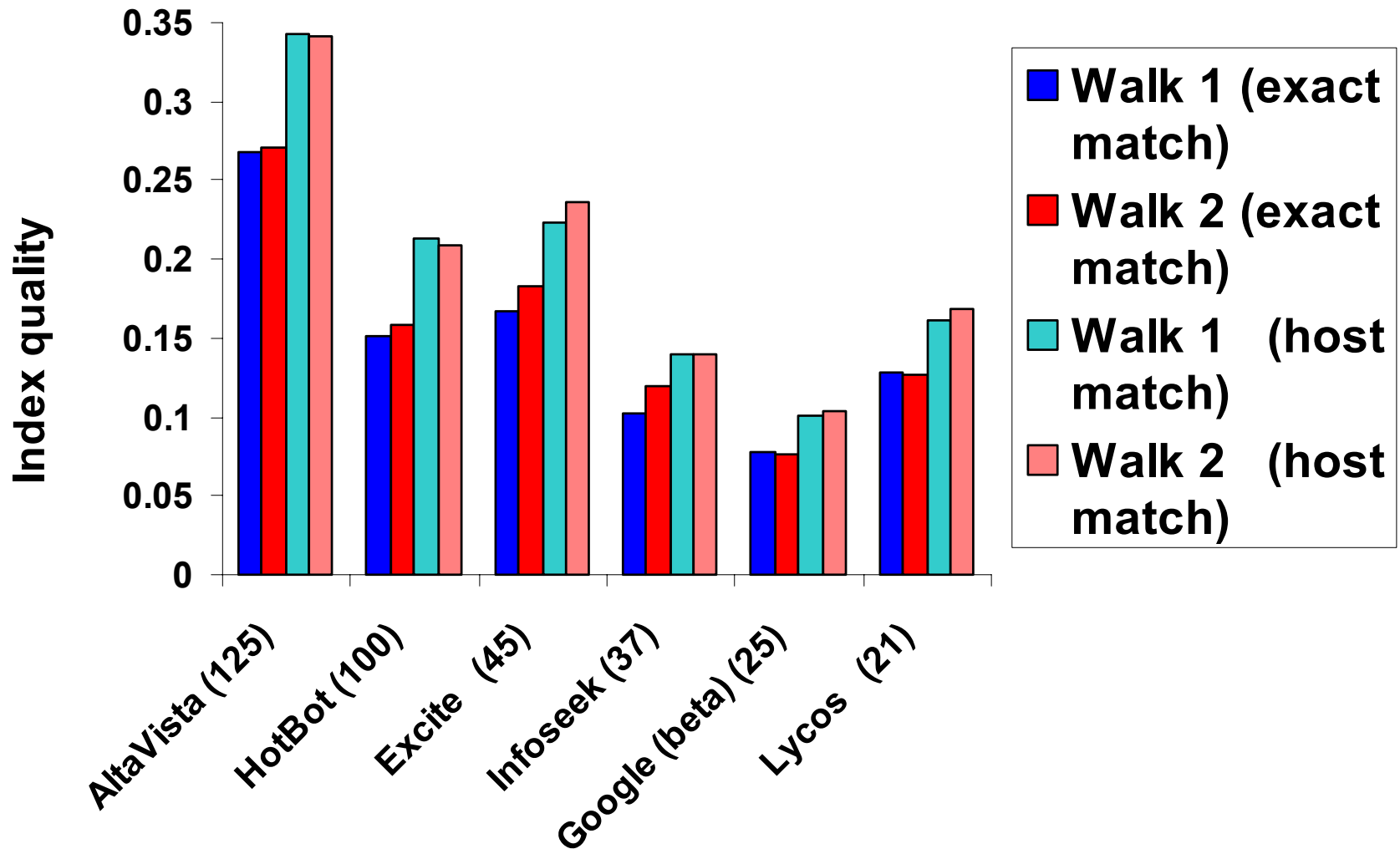
---

- Choose a sample of pages  $p_1, p_2, p_3 \dots p_n$  according to PageRank distribution  $w$
- Check if the pages are in search engine index  $S$  [BB'98]:
  - Exact match
  - Host match
- **Estimate for quality of index  $S$**  is the percentage of sampled pages that are in  $S$ , i.e.

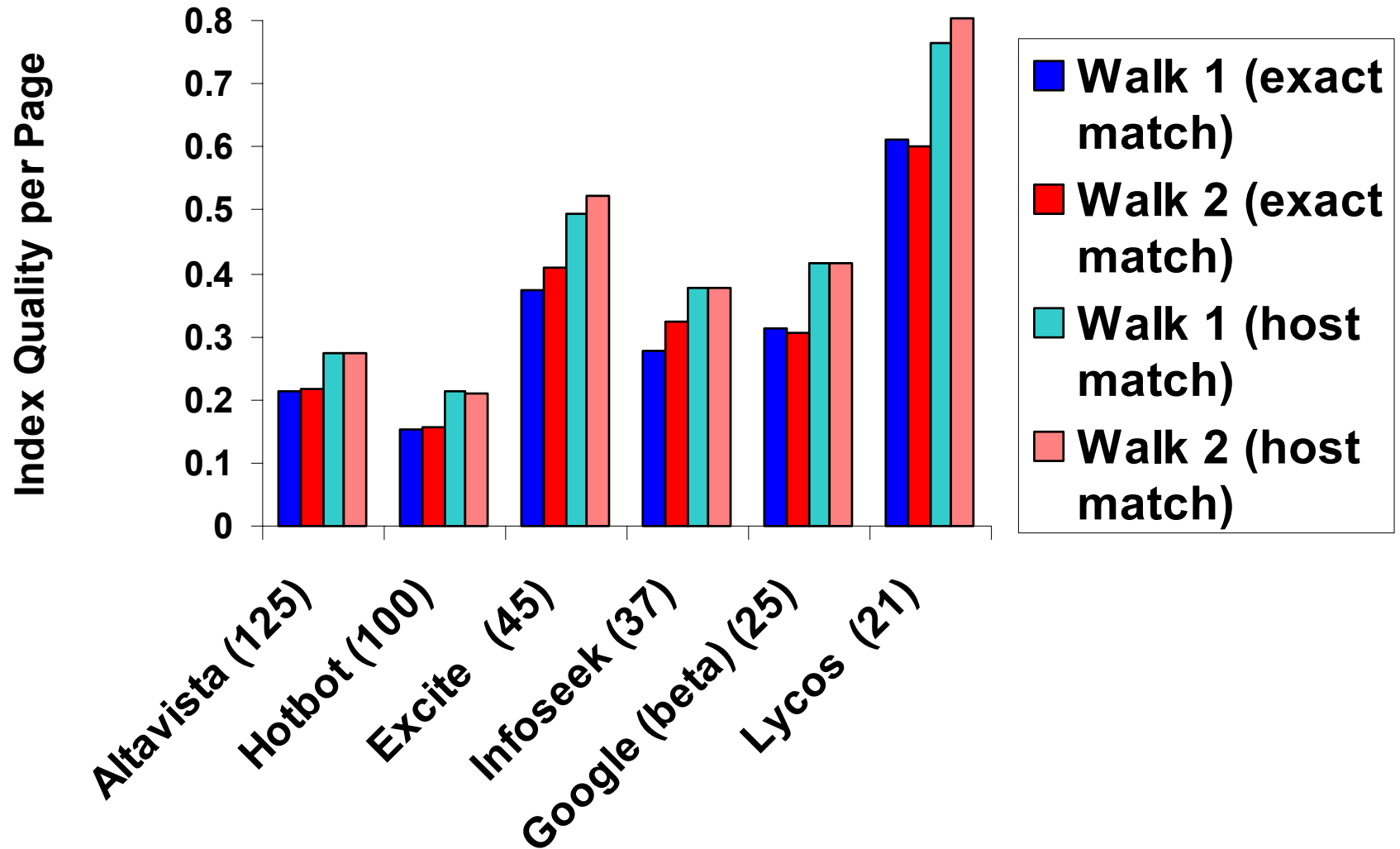
$$\bar{w}(S) = \frac{1}{n} \sum_j I[p_j \in S]$$

where  $I[p_j \text{ in } S] = 1$  if  $p_j$  is in  $S$  and 0 otherwise

# Results for index quality (fall'98)



# Results for index quality/page (fall '98)



# Sampling pages nearly uniformly

---

- Perform PageRank random walk
- Sample pages from walk s.t.

$$\Pr(p \text{ is sampled} \mid p \text{ is crawled}) \propto 1 / \text{PageRank}(p)$$

- “Nearly uniform” sample of the web
- Useful for estimation:
  - Web properties: Percentage of pages in a domain, in a language, on a topic, ...
  - Search engine comparison: Percentage of pages in a search engine index (**index size**)

# Sampling pages nearly uniformly

---

- “Nearly uniform” sample:

$$\Pr(p \text{ is sampled}) = \Pr(p \text{ is crawled}) \cdot \Pr(p \text{ is sampled} \mid p \text{ is crawled})$$

- A page is *well-connected* if it can be reached by almost every other page by *short* paths ( $O(n^{1/2})$  steps)
- For short paths in a well-connected graph:

$$\begin{aligned}\Pr(p \text{ is crawled}) &\approx E(\text{number of visits of } p) \\ &\approx L \cdot \text{PageRank}(p)\end{aligned}$$

# Sampling pages nearly uniformly

---

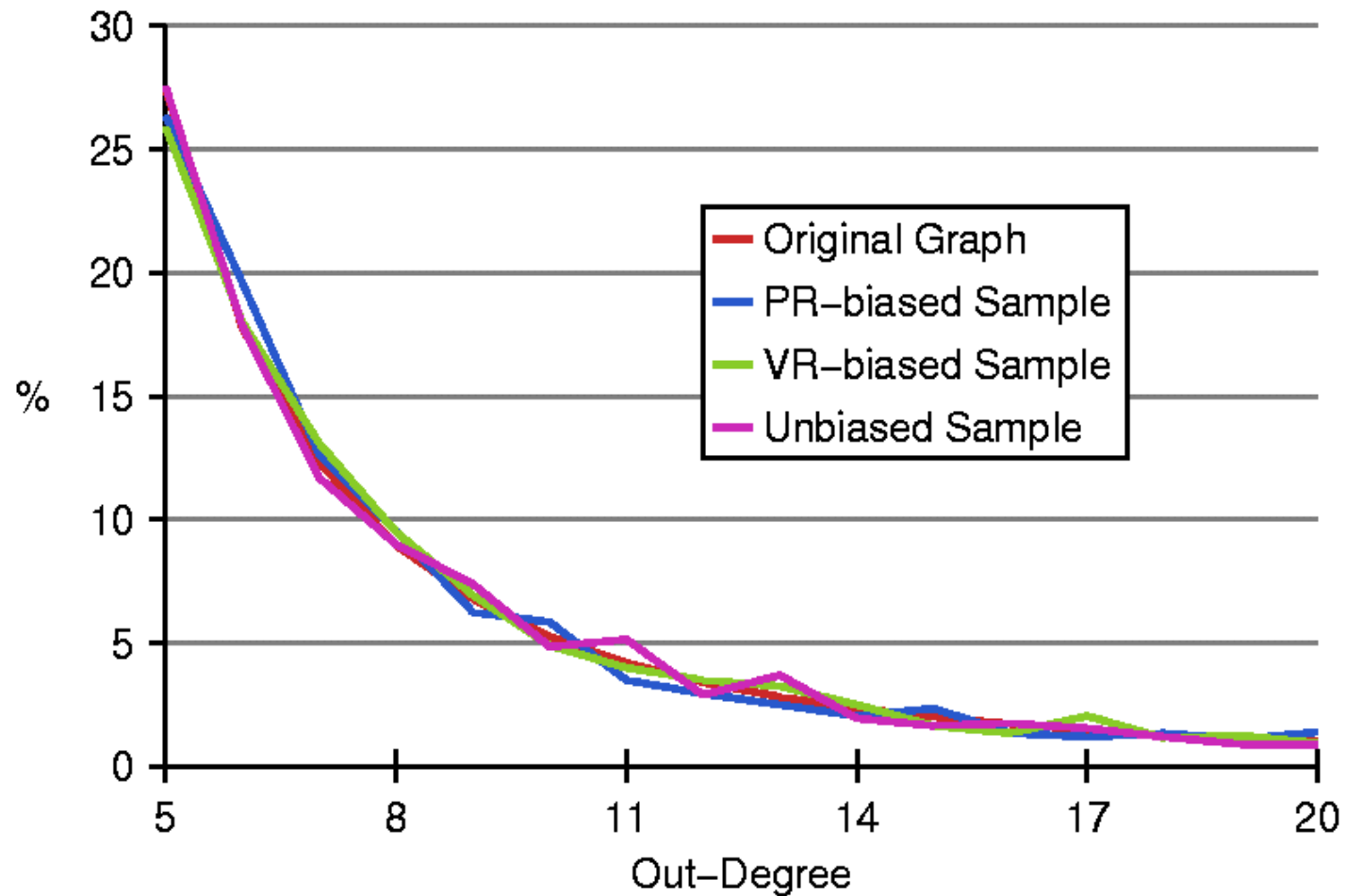
- Problems:
  - All previous problems
  - Need to approximate PageRank:
    - PR: PageRank computation of crawled graph
    - VR: VisitRatio on crawled graph
  - Dependence, especially in short cycles

# Evaluation using synthetic graph

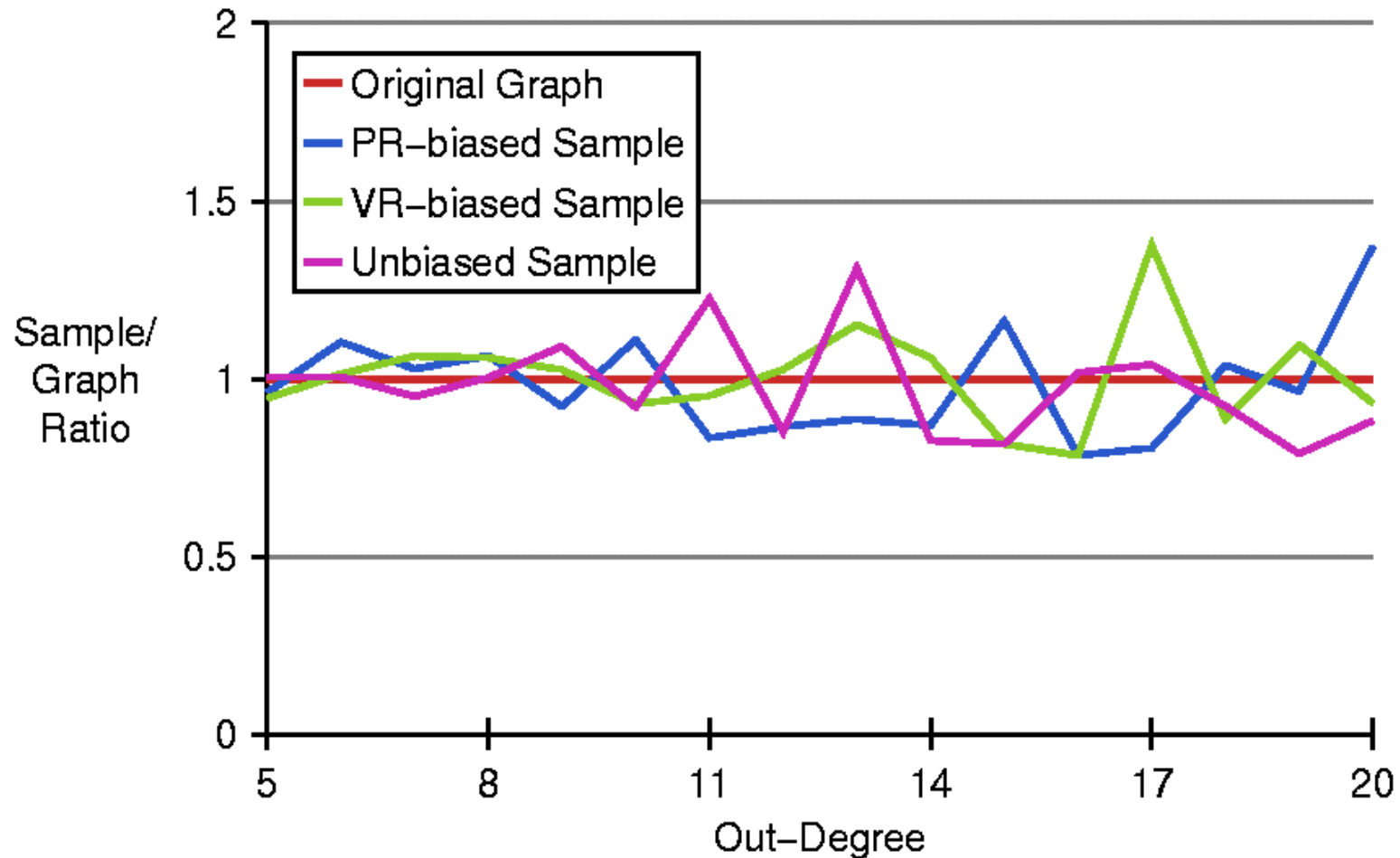
---

- Generated graph that mimics connectivity of real web (Zipfian distribution of in- & out-degree)
- Performed near uniform sampling using both PR and VR
- Compared connectivity characteristics of sampled nodes to those of entire graph
- If sampling were truly uniform, characteristics should be identical

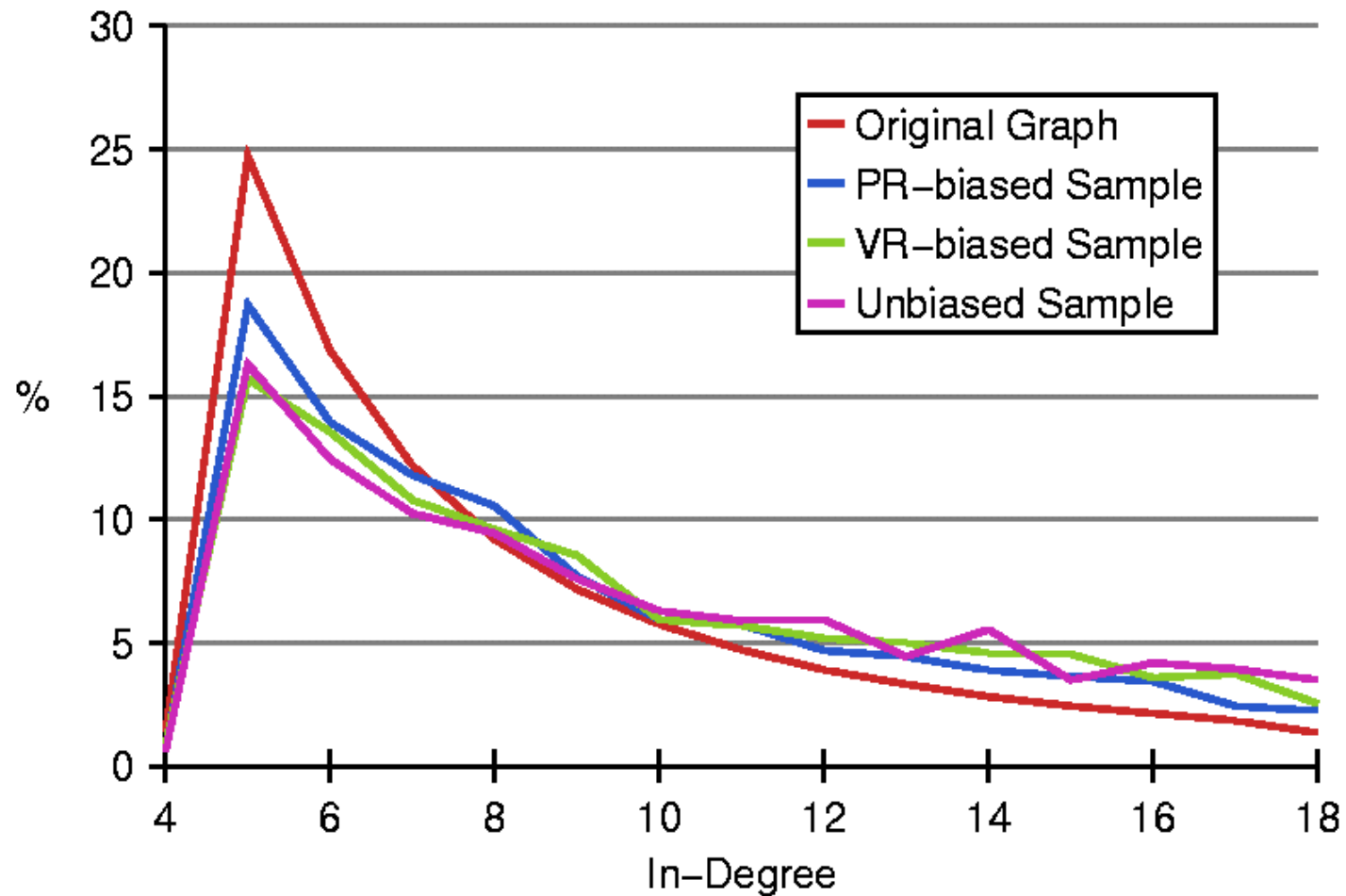
# Evaluation based on out-degree



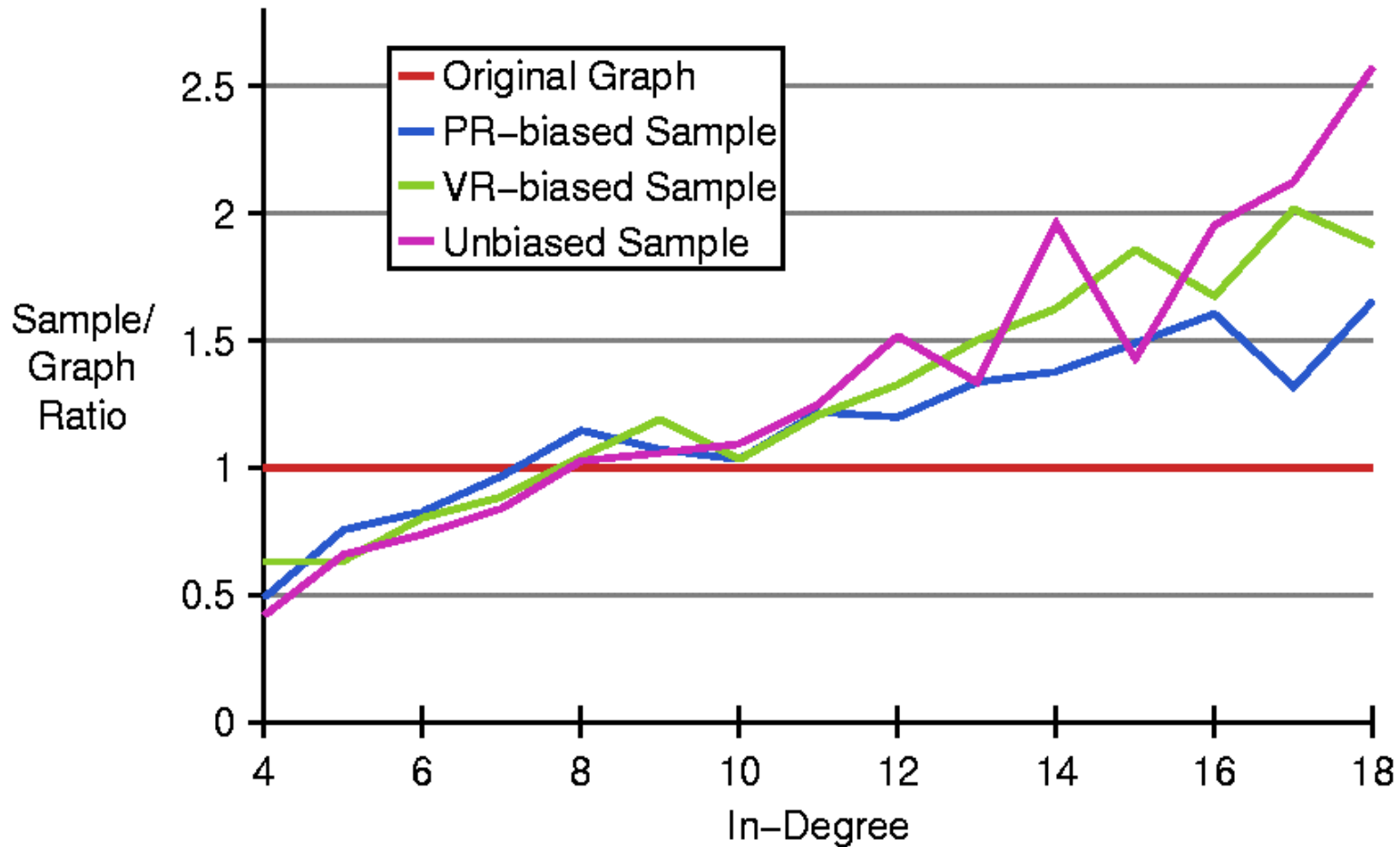
# Evaluation based on out-degree



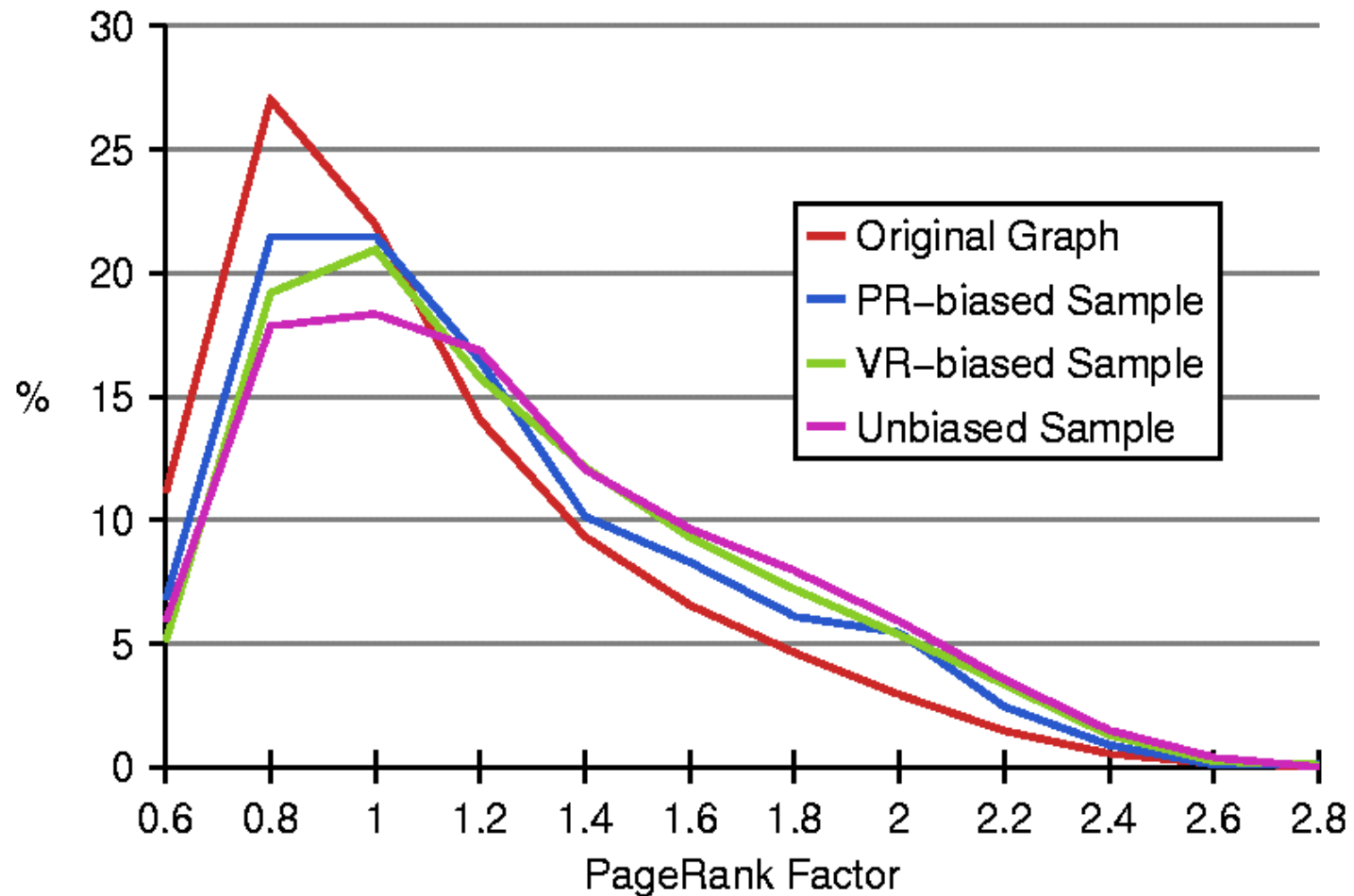
# Evaluation based on in-degree



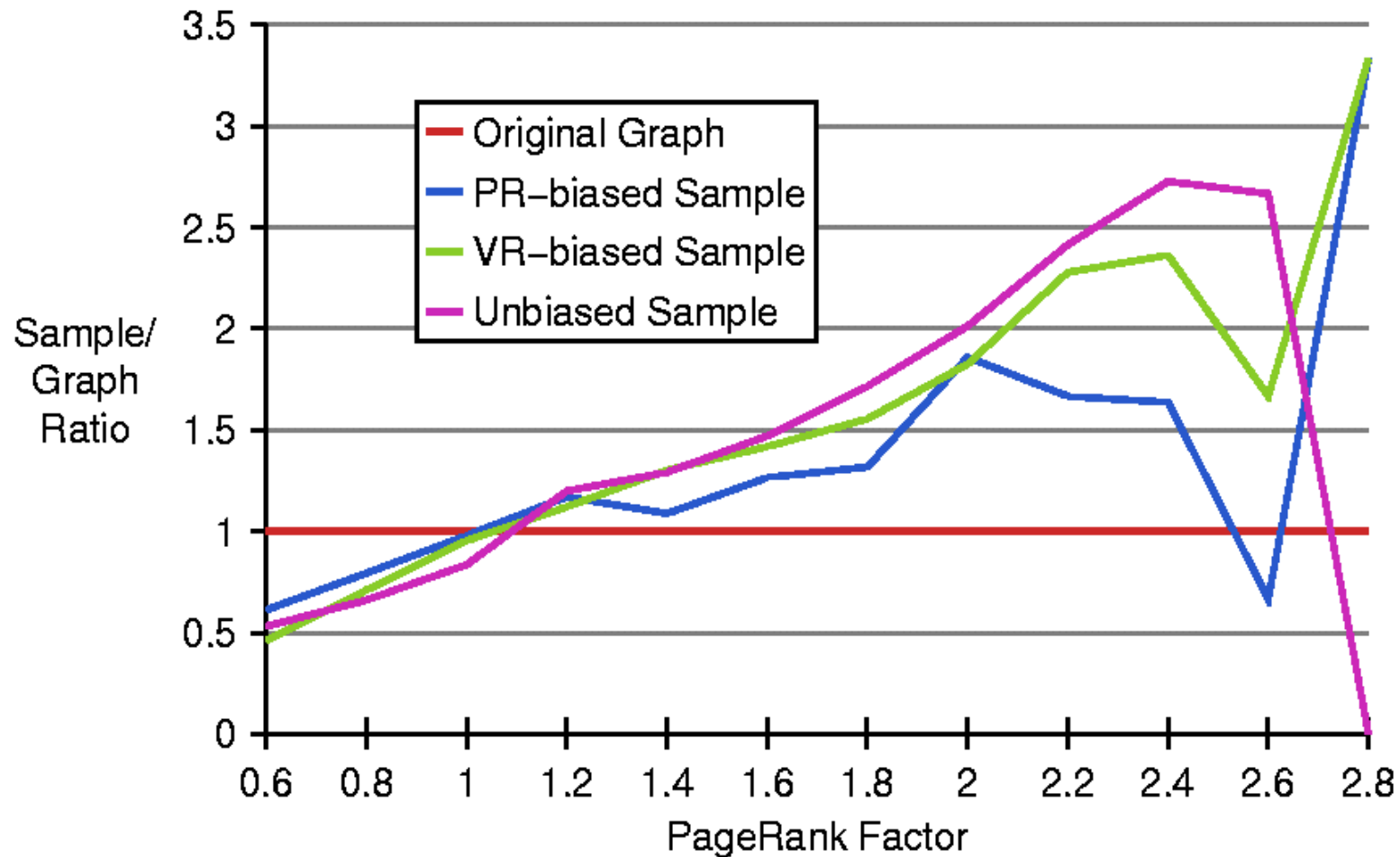
# Evaluation based on in-degree



# Evaluation based on PageRank

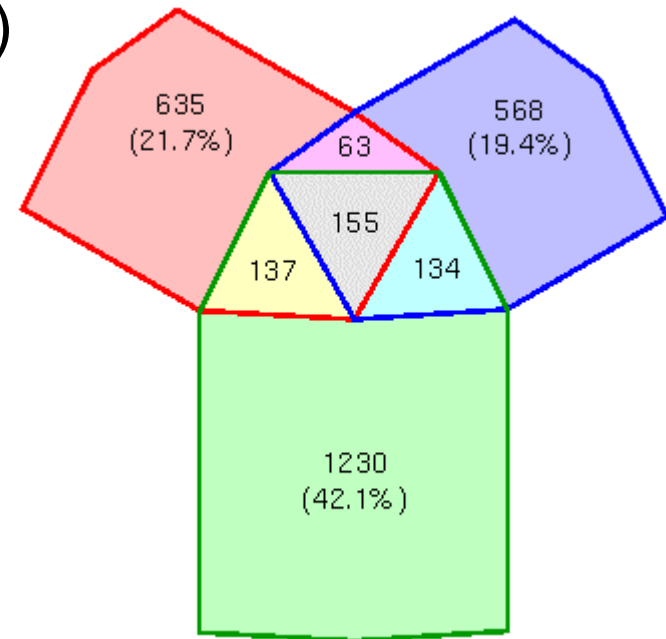


# Evaluation based on PageRank



# Experiments on the real web

- Performed 3 random walks in Nov 1999 (starting from 10,258 seed URLs)
- Small overlap between walks – walks disperse well (82% visited by only 1 walk)



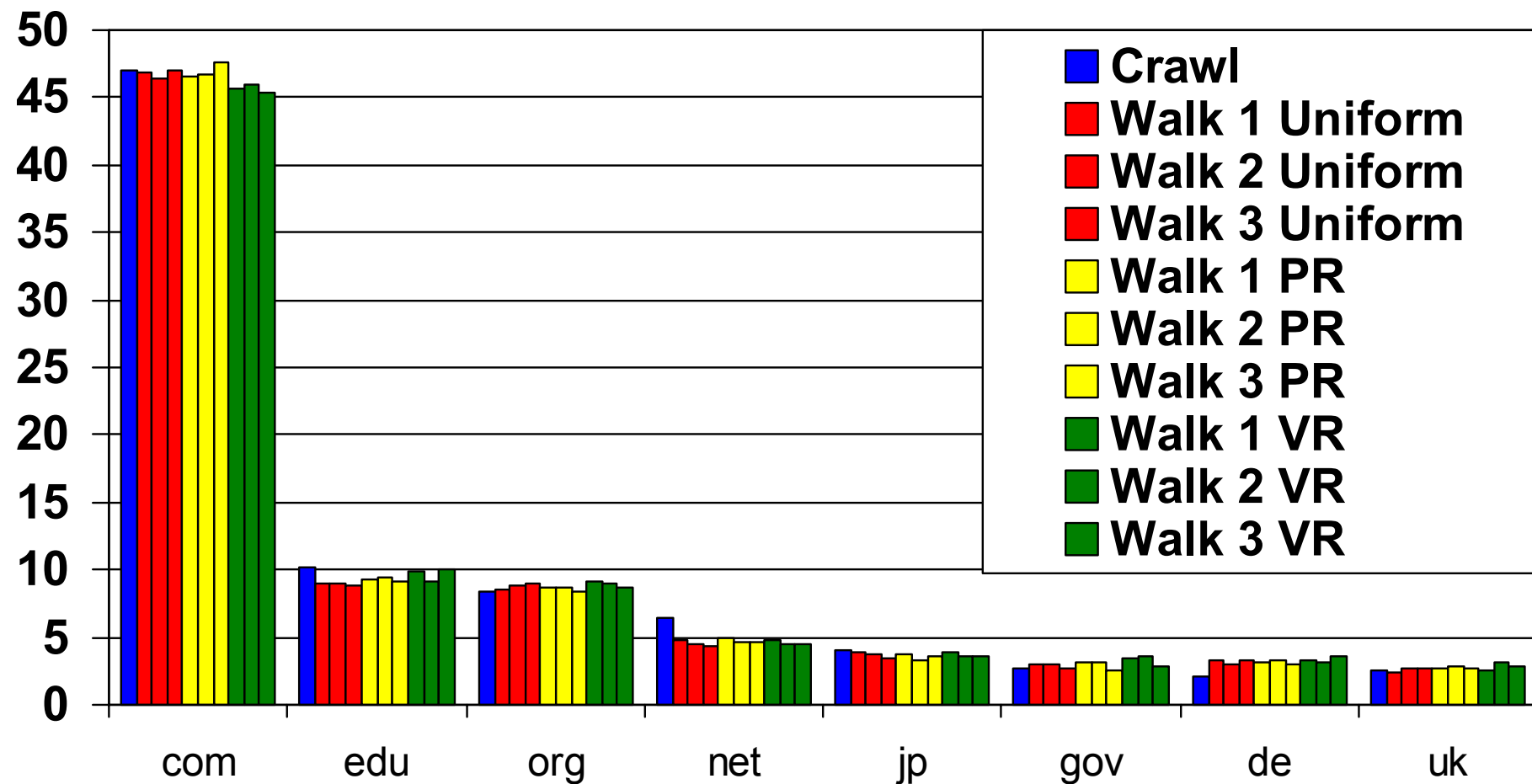
<u>Walk#</u>	<u>visited URLs</u>	<u>unique URLs</u>
1	2,702,939	990,251
2	2,507,004	921,114
3	5,006,745	1,655,799

# Experiments on the real web (cont.)

---

- Sampled each walk:
  - Uniform sampling
  - VR sampling
  - PR samplingTotal of 9 samples, each containing 10,000 URLs
- 2 Experiments:
  - Computed distribution of top-level domains of URLs in each sample and compared to distribution discovered during an 80m document web crawl
  - Index size comparison on 8 search engines

# Percentage of pages in domains



# Estimating search engine index size

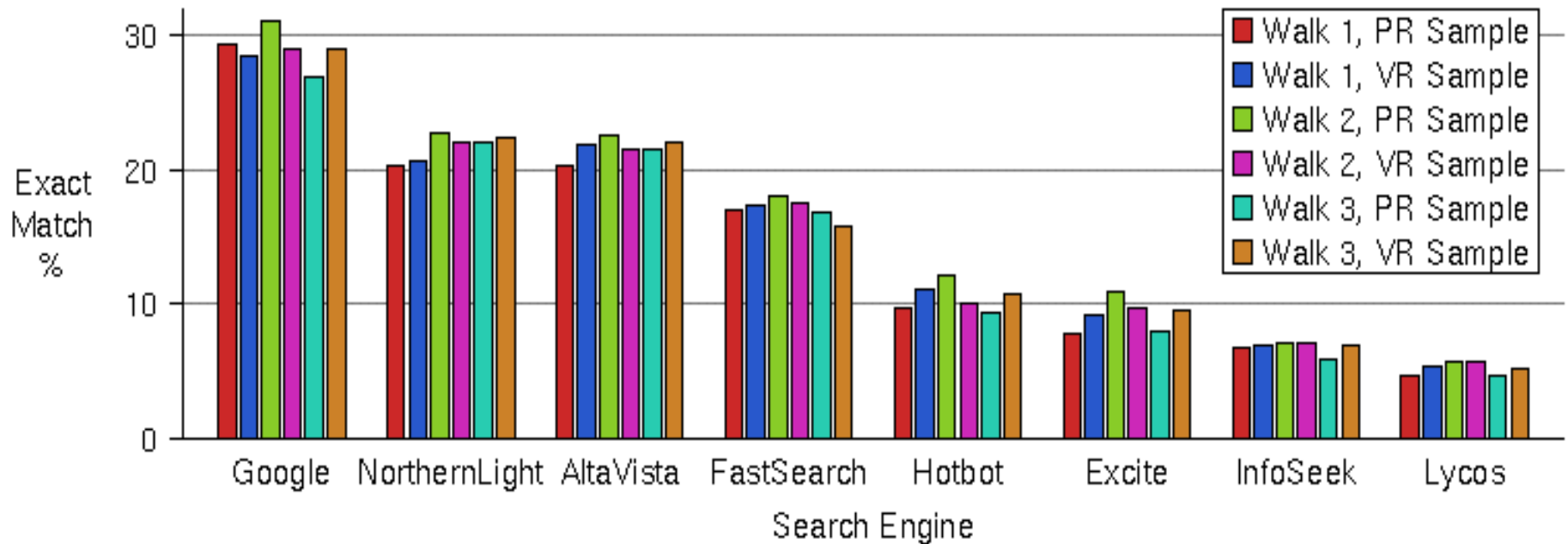
---

- Choose a sample of pages  $p_1, p_2, p_3 \dots p_n$  according to **near uniform** distribution
- Check if the pages are in search engine index  $S$  [BB'98]:
  - Exact match
  - Host match
- **Estimate for size of index  $S$**  is the percentage of sampled pages that are in  $S$ , i.e.

$$\bar{v}(S) = \frac{1}{n} \sum_j I[p_j \in S]$$

where  $I[p_j \text{ in } S] = 1$  if  $p_j$  is in  $S$  and 0 otherwise

# Result set for index size (fall'99)



# Summary

---

- Our random walks over-sample well-connected pages
- We compensate by sampling pages visited during random walk such that well-connected pages are less likely to be sampled
- Resulting sample is less skewed than random walk, but still not uniform

# Other approaches

---

- Lawrence and Giles '99
- Bar-Yossef et al '00
- Rusmevichientong et al '01