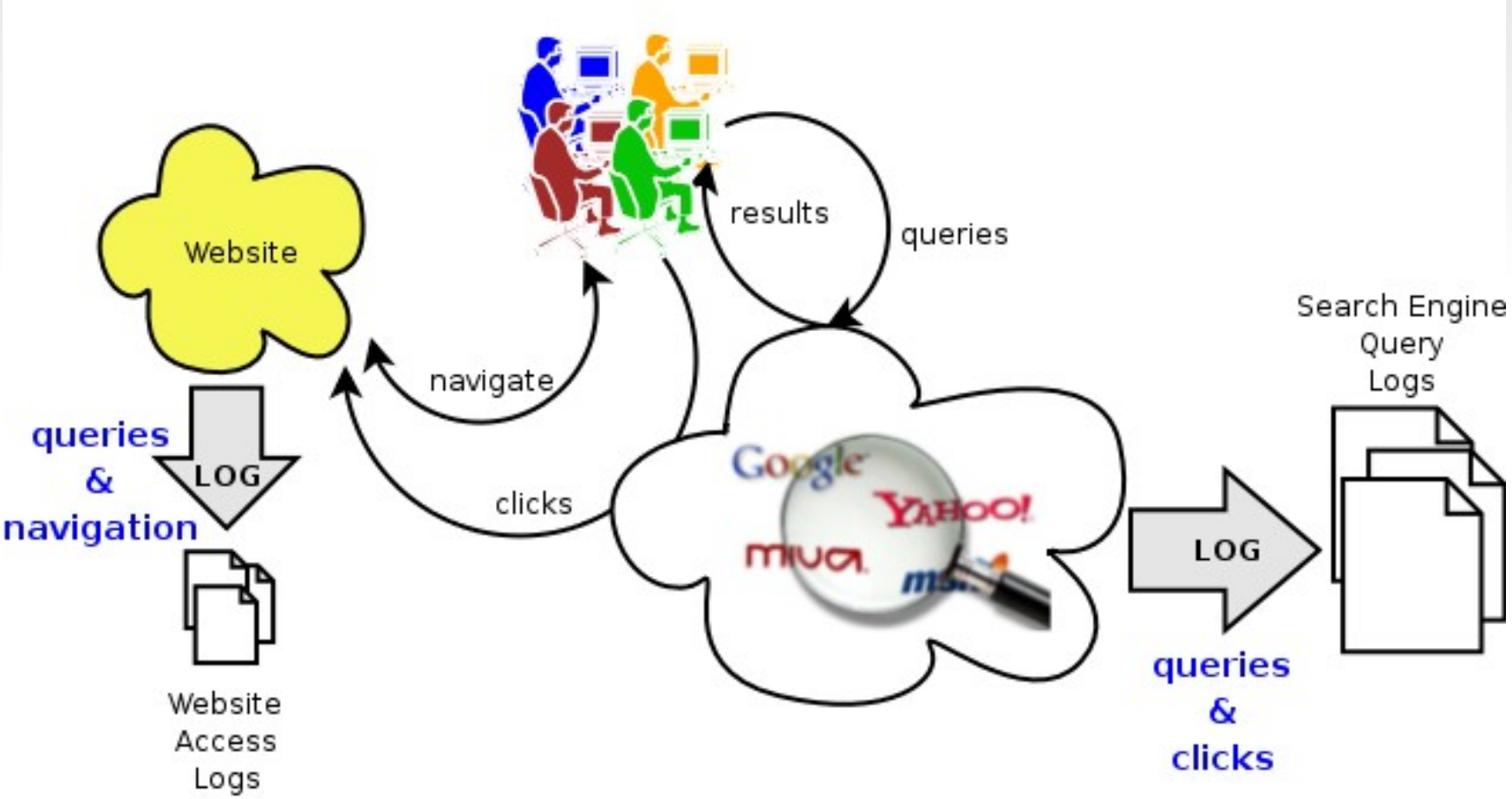
Website Privacy Preservation for Query Log Publishing

Barbara Poblete[1,2],
Myra Spiliopoulou[3] and Ricardo Baeza-Yates[2]

[1] University Pompeu Fabra, Barcelona, Spain.
[2] Yahoo! Research Barcelona, Spain.
[3] Otto-von-Guericke-University, Magdeburg, Germany.
Contact: barbara.poblete@upf.edu

Introduction

- **Query logs** are **very rich** sources of information.
  - They allow to discover interesting **behaviour patterns**.



Introduction…

- Some applications of usage logs (**query logs + access logs**):

  - Search engine ranking improvement, e.g.: re-ordering of results based on user feedback.
  - Search engine results personalization.

  Search Engine Query Logs

  - Adaptive websites.
  - Personalization.
  - General website improvement (off-line, and not personalized):
    - based on navigational patterns.
    - based on navigation and user queries.

  Website Access Logs

Introduction…

- Owners of query logs do **not publish their data**, due to important **privacy concerns**.
  - Triggered by the publication of the **AOL** log (**20M queries** from **650k users**).
  - Only **naive anonymization** was performed on it.
  - Many users issued **self identifying** (or semi-identifying) **queries**, which made it possible to map them to real people.

Introduction…

« Previous post Next post »

AOL Proudly Releases

The New York Times **Technology**

A sample of Thelma Arnold's search data released by AOL

4417749	swing sets	2006-04-24	15:39:30	4	http://www.byoswingset.com
4417749	swing sets	2006-04-24	15:39:30	9	http://www.buychoice.com
4417749	swing sets	2006-04-24	15:39:30	10	http://www.creativeplaythings.com
4417749	swing sets	2006-04-24	15:39:30	5	http://www.childlife.com
4417749	swing sets	2006-04-24	15:39:30	6	http://www.planitplay.com
4417749	that do not shed	2006-04-28	9:05:54	2	http://www.gopetsamerica.com
4417749	dog who urinate on everything	2006-04-28	13:24:07	6	http://www.dogdaysusa.com
4417749	walmart	2006-04-28	14:07:32	1	http://www.walmart.com
4417749	womens underwear	2006-04-28	14:12:28	10	http://www.bizrate.com
4417749	jcpenny	2006-04-28	14:16:05		
4417749	jcpenny	2006-04-28	14:16:49	1	http://www.jcpenney.com
4417749	tortus and turtles	2006-04-29	13:12:47		
4417749	manchester terrier	2006-05-02	9:05:31	1	http://www.manchesterterrier.com
4417749	delta	2006-05-02	11:49:26		
4417749	fingers going numb	2006-05-02	17:35:47		
4417749	dances by laura	2006-05-02	17:59:32		
4417749	dances by lori	2006-05-02	17:59:57		
4417749	single dances	2006-05-02	18:00:18	1	http://solosingles.com
4417749	single dances in atlanta	2006-05-02	18:01:13		
4417749	single dances in atlanta	2006-05-02	18:01:50		
4417749	dry mouth	2006-05-06	16:49:14	2	http://www.mayoclinic.com
4417749	dry mouth	2006-05-06	16:49:14	8	http://www.wrongdiagnosis.com
4417749	tyroid	2006-05-06	16:55:34		
4417749	thyroid	2006-05-06	16:55:44		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:14:52		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:16:17		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:16:43		

Why the search

"I was thinking about my grandchildren"

"I was looking for some."

"A woman was in the [public] bathroom crying. She was going though a divorce. I thought there was a place called 'Dances by Lori,' for singles."

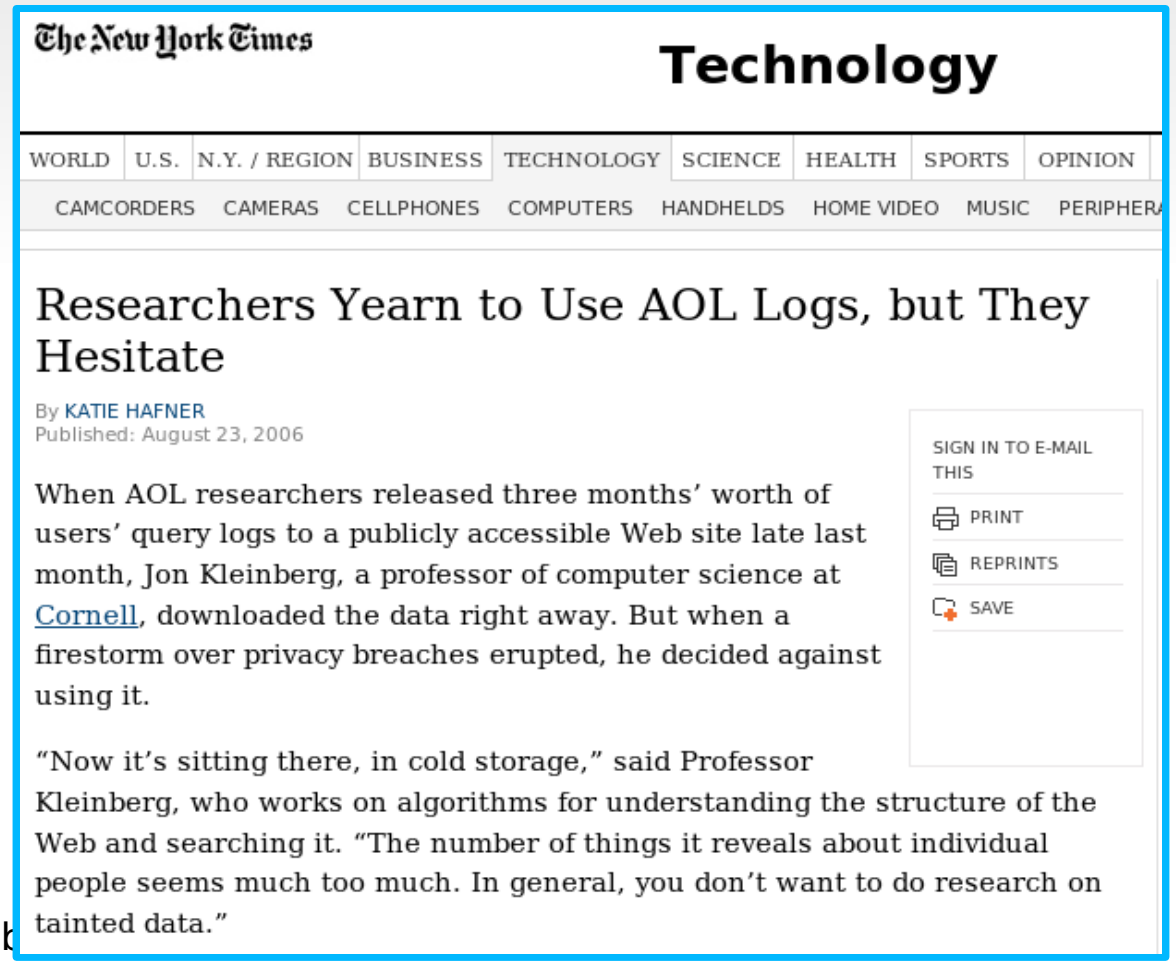
"I wanted to find out what my house was worth."

their permission. While the AOL username ha
analyze all searches by a single user will ofte
what they are up to. The data includes perso
everything else someone might type into a s

AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

Introduction…

- This **created awareness** among users (also researchers) and providers.
- Very **unlikely** that other releases are made until more guaranties are given.



The New York Times

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

CAMCORDERS | CAMERAS | CELLPHONES | COMPUTERS | HANDHELDS | HOME VIDEO | MUSIC | PERIPHERA

Researchers Yearn to Use AOL Logs, but They Hesitate

By KATIE HAFNER
Published: August 23, 2006

SIGN IN TO E-MAIL THIS
PRINT
REPRINTS
SAVE

When AOL researchers released three months' worth of users' query logs to a publicly accessible Web site late last month, Jon Kleinberg, a professor of computer science at Cornell, downloaded the data right away. But when a firestorm over privacy breaches erupted, he decided against using it.

"Now it's sitting there, in cold storage," said Professor Kleinberg, who works on algorithms for understanding the structure of the Web and searching it. "The number of things it reveals about individual people seems much too much. In general, you don't want to do research on tainted data."

Introduction...

- Privacy preservation in query logs is a **new scientific matter**.
- Research is needed to identify and tackle **all possible risks** involved in publishing a query log.
- There are two specially revealing fields registered in a query log:
  - the **query**, and
  - the clicked **URLs** for that query.
- This information can be revealing not only from the **user** point of view, but also revealing of **website** (or **business**) confidential information.
- Search engines rely on the relationship with their users and with content providers.
- A search engine **should not compromise the security or confidentiality of a website's users or contents**.

Introduction…

- Sensitive website information which can be discovered from a query log, this includes:
  - Most important queries for a website (placement of competitor ads).
  - Traffic, i.e.: ~ 80% of traffic to websites can come from search engines
    ⇒ using a log of an important search engine one could recreate the access log of a site.
  - Visits to each document in the site and queries used to reach them.
  - Measure a company's performance based on queries and the documents that were visited. Find successful queries (strengths) and unsuccessful queries (weaknesses) on its own, or in relation to other competitors.
  - Business decisions could be made based on this type of knowledge.
  - These are only a FEW examples.

Outline

- We introduce a **new privacy concern**: **website privacy**.
- We show **attacks that disclose confidential information** about a website, and ways to prevent them.
- We propose a **graph-based method** for log anonymization.
- We show a first **experimental analysis over real data** to validate our heuristic.

Scope of Our Work

- We focus on website privacy preservation in the **context of publishing query logs of search engines**.
- We separate this problem from that of user privacy preservation.
- Our goal is to **prevent an adversary from obtaining confidential information** about traffic to websites.

Why Website Privacy Preservation?

- A website presents many services/products offered by a company.
- The website traffic reflects the **interaction of the company with its customers** and potential customers. **This information is CONFIDENTIAL**
- An **important fraction of website traffic comes from search engines**.
- A query log can be **(mis)used to discover confidential information** about a website:
  - Patterns that are more informative than a **market study**.
  - Information exploitable for **industrial espionage** and **other malicious activities.**

Query Log Signature (based on the AOL release):

`UserId`, `Query`, `Timestamp`, `ItemRank`, `ClickedURL`

Why is Website Privacy Preservation Difficult?

- An **adversary** that attacks a website **W** can **combine information** from **public and private data sources**.
  - The adversary may be a **competitor of W** and have an **own website W'**:
    - The adversary has access to all of the traffic data for W' including its queries (**access log W'**). **PRIVATE DATA**
    - The adversary can have access to the **query log of W** (extracted from the published query log, even if anonymized). **PUBLIC DATA**
    - The adversary can **detect popular queries** that reach **both W' and W**.
    - The adversary has **background knowledge on the market**, including user (customer) profiles and patterns of user behaviour.

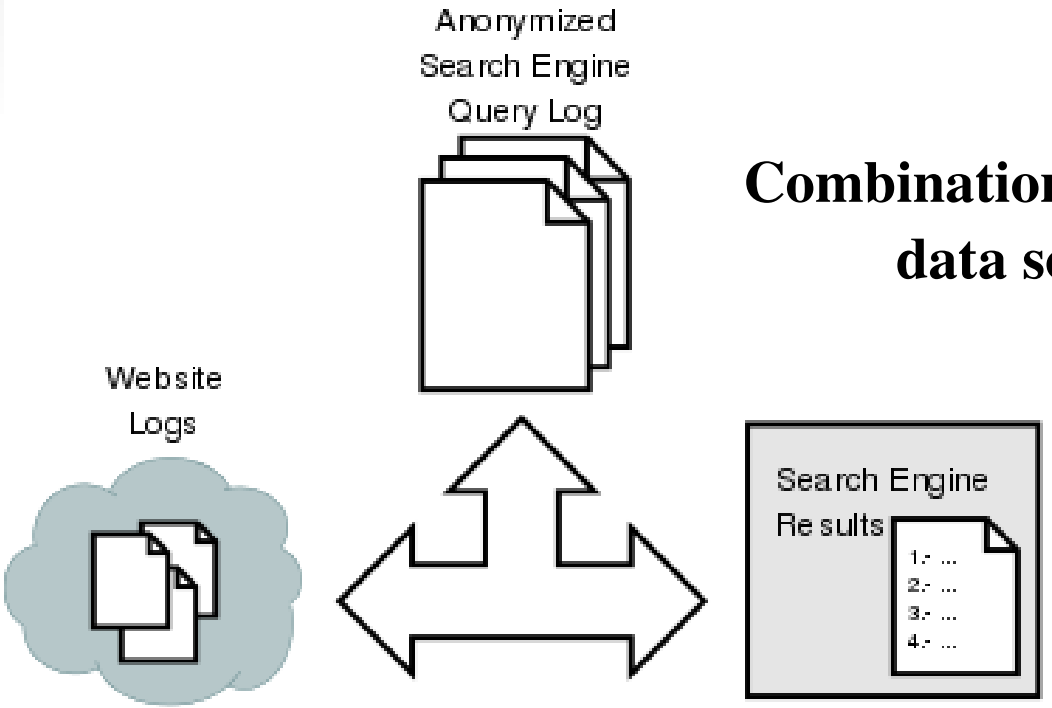
Website privacy breach:
An adversary collects public information from the website, possibly combines it with already owned information, and concludes confidential information about the website's content.

The Data Sources

1. Website log, containing also requests that come from a search engine.
2. Anonymized query log of a search engine.
3. Search results for queries launched to a search engine.

Private

Public



Combination of multiple data sources

Related Work

- *k*-anonymity [P. Samarati & L. Sweeney, 1998]
  - *k*-anonymity is a powerful concept.
  - According to [E. Adar, WWW'07], enforcing *k*-anonymity upon rapidly changing data (e.g. query logs) is too expensive.
    Two specific solutions are proposed for *user anonymization*.
  - *k*-anonymity has been designed for a single data source.
    It cannot apply to data records built through the combination of data from multiple databases (using joins, incl. outerjoins)
- Query log anonymization [R. Kumar et. al, WWW'07] using token-based hashing.
- Ours is the **first work to address the issue of privacy preservation in query logs for websites** rather than for users.

A Few Words on Query Log Anonymization

- **Problem:**
  Which attributes should be anonymized so that the private information that can be disclosed is minimized while keeping the capability of mining applications?
- **Facts:**
  - The **exact values** of each attribute **can be hidden** behind an anonymized value that **preserves their distribution**.
  - **All attributes are potentially useful**.
  - Attributes are **not independent**, e.g.
    - Frequency of terms and co-occurrences can be exploited to compromise anonymization [R. Kumar et. al, WWW'07].
  - The **order of the query records is important**, records may not be shuffled as part of the anonymization process.

Adversaries – Two types

- **General Adversary:**
  An agent to whom any information is of potential interest.
  - Institutions that perform market studies.
  - Companies that provide the service of
    "improving the ranking of a website in a search engine" a.k.a. S.E.O.
- **Adversarial Competitor:**
  An agent that is interested in information about given websites.
  - Institution that collects information about competitors:
    It has background knowledge about the market and owns also a website.

Outline

- ✓ We introduce a **new privacy concern**: **website privacy**.
- We show **attacks that disclose confidential information** about a website, and ways to prevent them.
- We propose a **graph-based method** for log anonymization.
- We show a first **experimental analysis over real data** to validate our heuristic.

Outline

- ✓ We introduce a **new privacy concern**: **website privacy**.
- We show **attacks that disclose confidential information** about a website,
  a) attacks based on *vulnerable* queries.
  b) attacks that exploit private information – the own website log.
  c) attacks that exploit knowledge about a particular user in the log.
- and ways to prevent them.
- We propose a **graph-based method** for log anonymization.
- We show a first **experimental analysis over real data** to validate our heuristic.

Query Log Anonymization

- AOL anonymization:

`AnonID, Query, Timestamp, ItemRank, SiteOfURL`

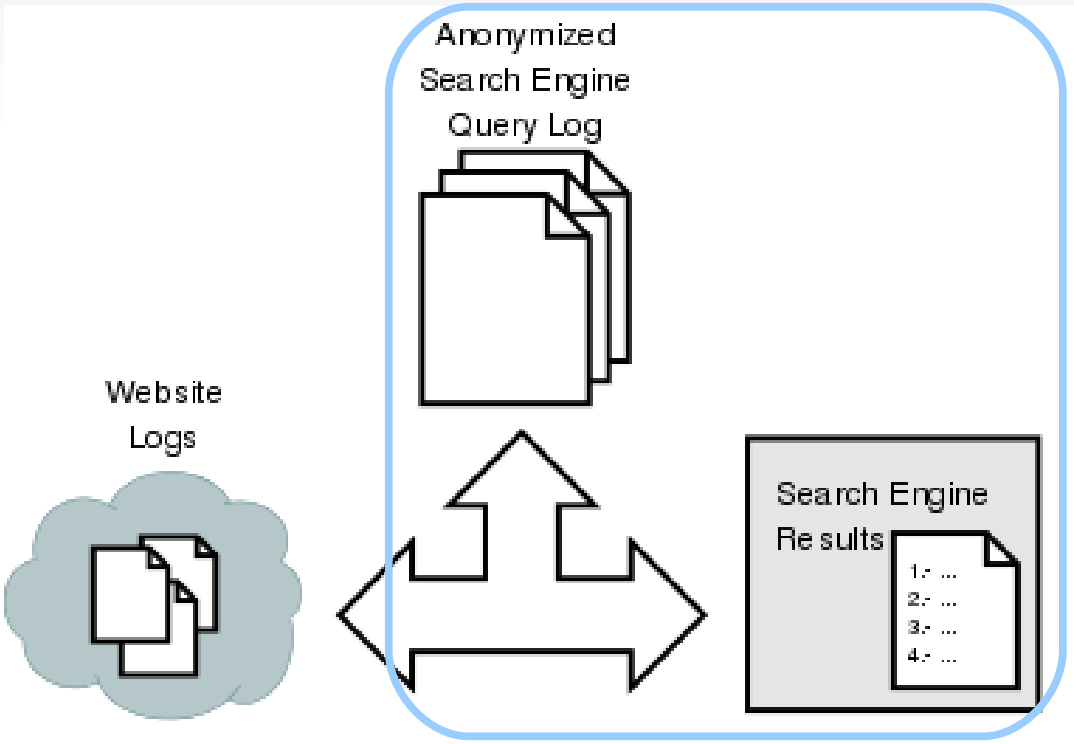
  - By issuing the **`Query`** to the search engine and by knowing the **`SiteOfURL`** and the **`ItemRank`**, one can deduce the complete **`ClickedURL`**.

Query Log Anonymization...

- Anonymization with **rank hiding**:
  - The **`ClickedURL`** is truncated at website level and the website is anonymized into **`AnonSiteID`**.
  - The **`ClickedURL`** is anonymized over the whole query log.
  - The **`ClickedURL`** is anonymized within the website, i.e. All URLs of the same site have the same **`AnonSiteID`**: **`AnonClickedURL = AnonSiteID++WithinSiteDocID`**

Attacks using Vulnerable Queries

1. Navigational queries.
2. Queries returning less than *k* results.
3. Groups of queries that share some of the clicked URLs.

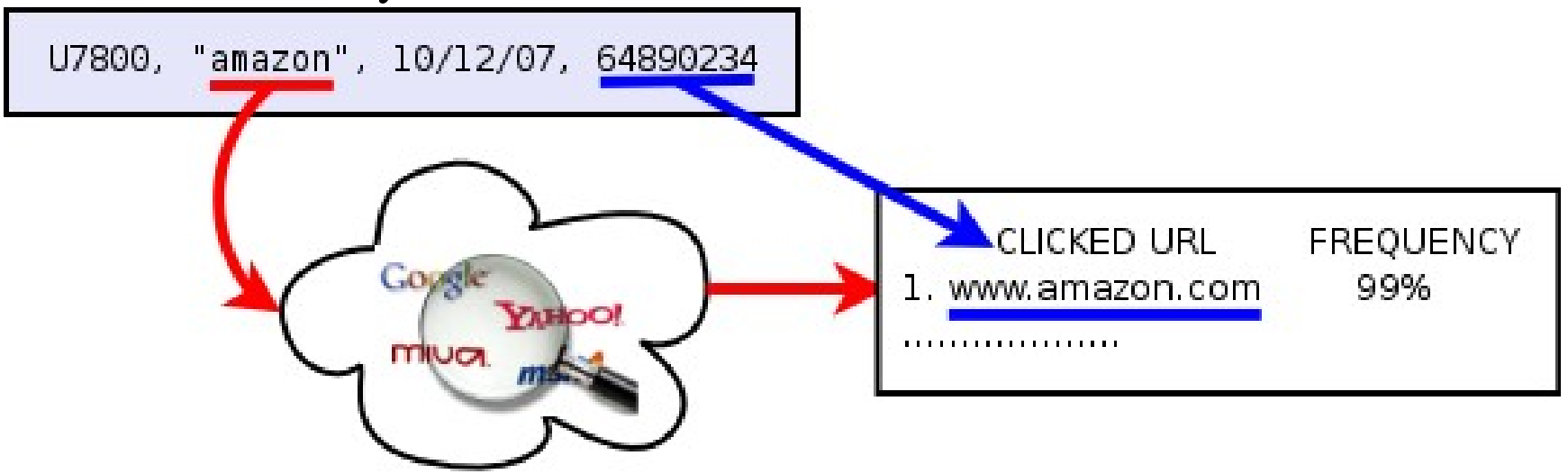


Attacks using Vulnerable Queries...

1. Very frequent **navigational** queries.
   Queries that contain part of the URL string of the targeted website, or for which there is a very strong association between the query and the clicked URL.

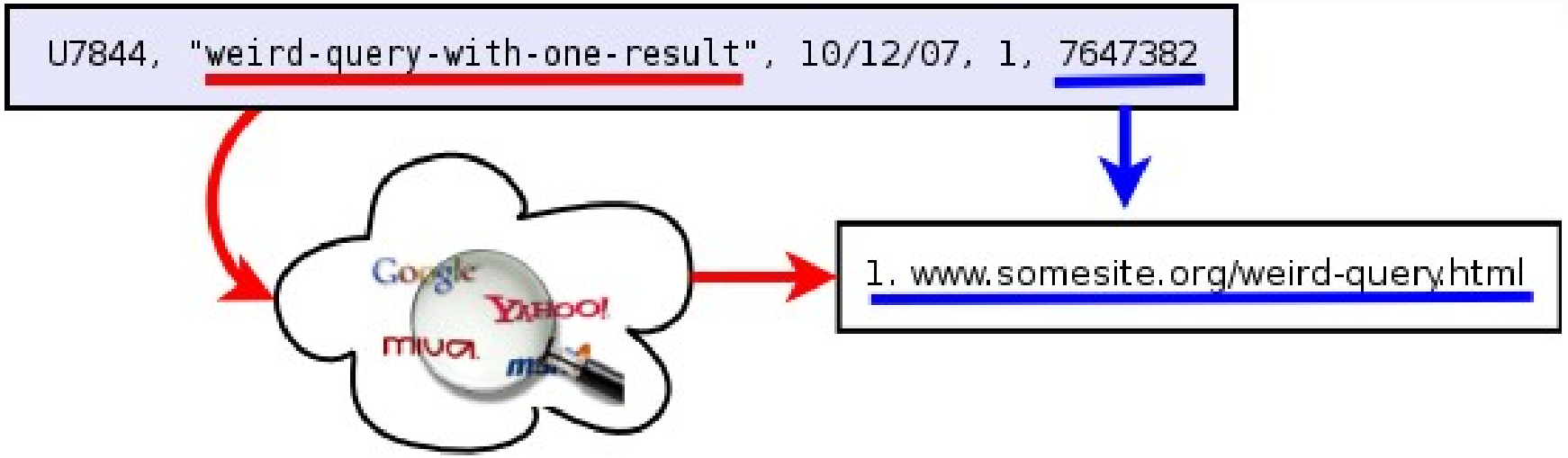
   ➔ These queries disclose the website URL.

   ➔ Additionally to the previous URL anonymization, this queries should be removed or anonymized.



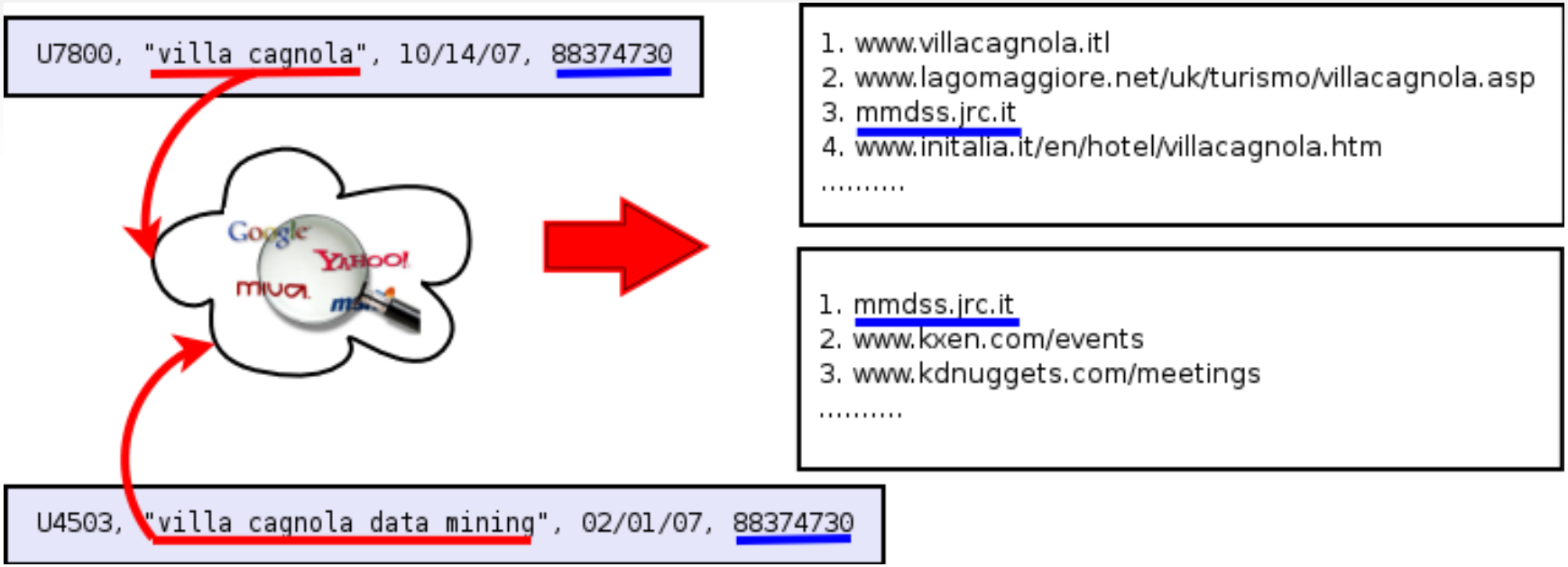
Attacks using Vulnerable Queries...

2. Queries returning less than k results
   - ➔ In the context of *k*-anonymity, the URLs returned by those queries can be disclosed.
   - ➔ Queries returning less than k results should be removed.



Attacks using Vulnerable Queries...

3. Groups of queries that share some of the clicked URLs
   ➔ They can disclose the URLs to an adversarial competitor.



Attacks using Vulnerable Queries...

3. Groups of queries that share some of the clicked URLs

Adversary defines group of queries Q_1.

Match Q_1 in anonymized log L and obtain set C of candidate anonymized URLs .

For each $u \in C$ collect from Q the queries that clicked on u into Q_u.

For each $q_i \in Q_u$ collect anonymized result set R_{Ai}.

For each pair of queries (q_i, q_j) with $|R_{Ai} \cap R_{Aj}| \geq 1$:

Build R_i, R_j from the search engine.

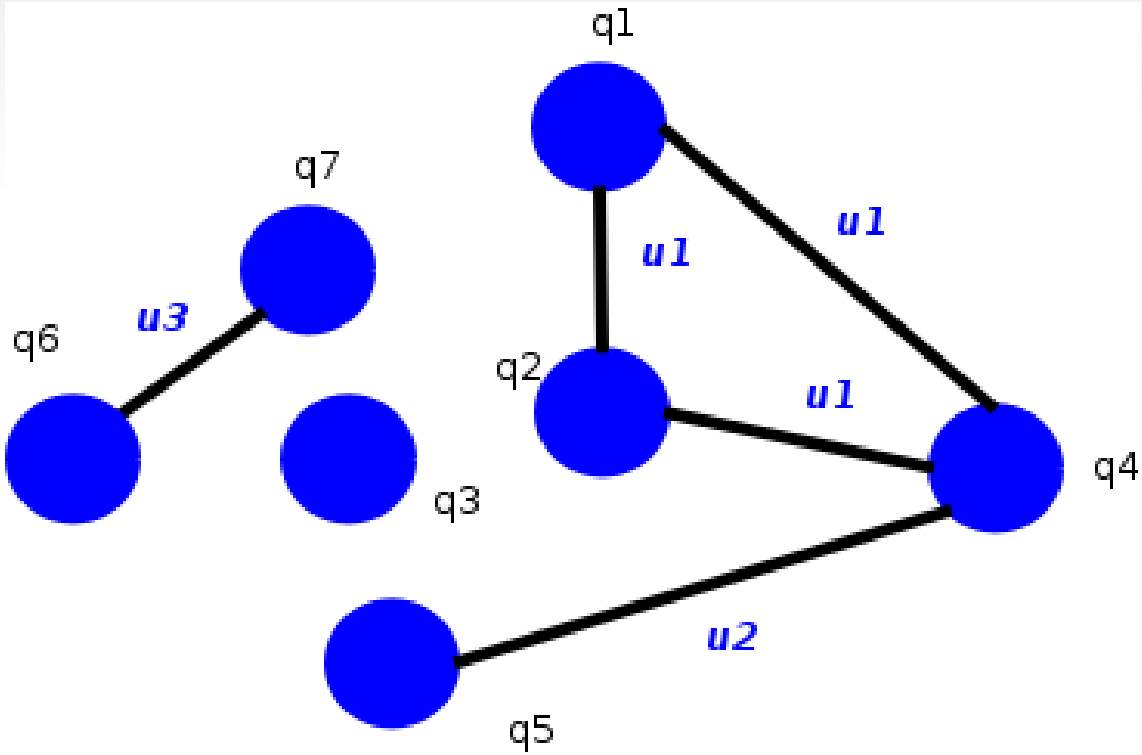
If $|R_{Ai} \cap R_{Aj}| = |R_i \cap R_j| = 1$ then exact match.

If $|R_{Ai} \cap R_{Aj}| = |R_i \cap R_j| > 1$ then approximate match, unless

all but one URL in $|R_{Ai} \cap R_{Aj}|$ have already been disclosed .

Heuristic Approach Against Attack 3

- Graph representation of the log:
  - A node is a query.
  - An edge connects two queries if there are URLs clicked by both of them.



Heuristic Approach Against Attack 3…

- Brute force approach for attack prevention:
  - Disconnect the graph by removing as few nodes as possible. (Maximum Independent Set, NP-Hard)
  - Greedy Heuristic:
    At each step, remove the node with the highest degree, until the graph is disconnected.
    1. Sort the nodes by in-degree.
    2. Remove the node with the highest degree.
    3. Recalculate the in-degrees of all the neighbor nodes.
    4. If the density is zero then stop, else iterate.
  - Greedy Heuristic+
    1. Sort the nodes by (in-degree/frequency)

Heuristic Approach Against Attack 3...

- NOTE:
  - The degree of a query is *not* reflecting the frequency of the query.
  - It turns out that queries with common results are rare.
  - In other words, the likelihood that two *frequent* queries have common results (and thus are connected) is low.
  - This is peculiar of this graph.
    But the graph is representative of query logs.

Other Attacks: Using Website Logs

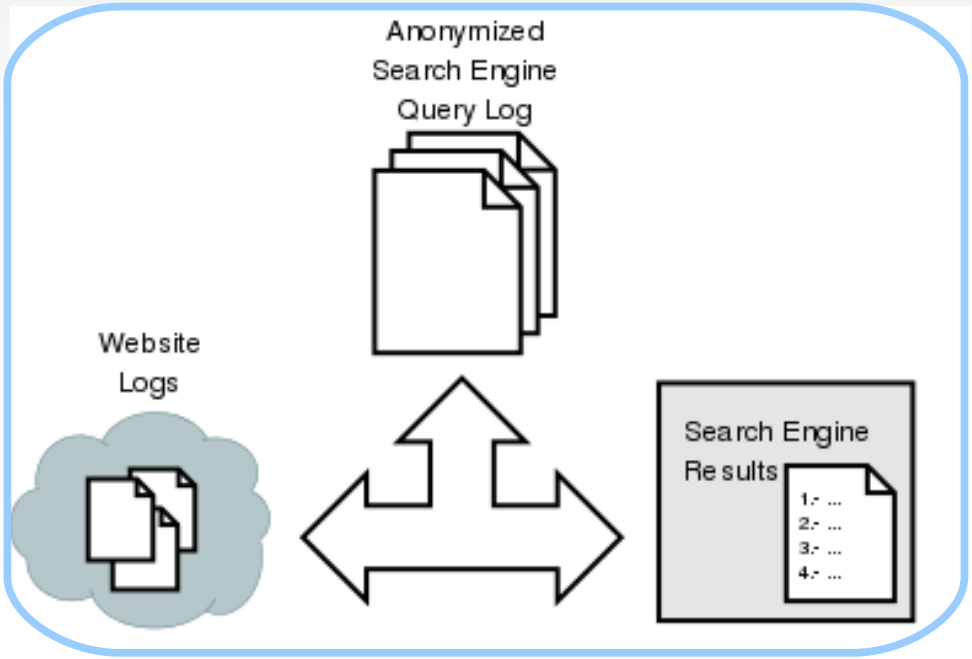
- A competitor has a website log that covers the same time period as the anonymized query log L.

  ➔ The adversary can match the own URLs to the anonymized URLs in L.

  - If adversaries collude, the combination of weblogs will disclose yet more URLs.

Solution:

  - If the results of two queries have a non-empty intersection, they must come from at least k different sites.



Other Attacks: User Identification

- The adversary acquires information about a user:
  - (the user's identity)
  - the user's search activities and
  - the search results – the clicked URLs
- The adversary traces the user in the anonymized log:
  - The adversary maps the anonymized URLs for this user to the previously disclosed clicked URLs.
  - ➔ The corresponding sites and further URLs of this site are disclosed.

Solution:

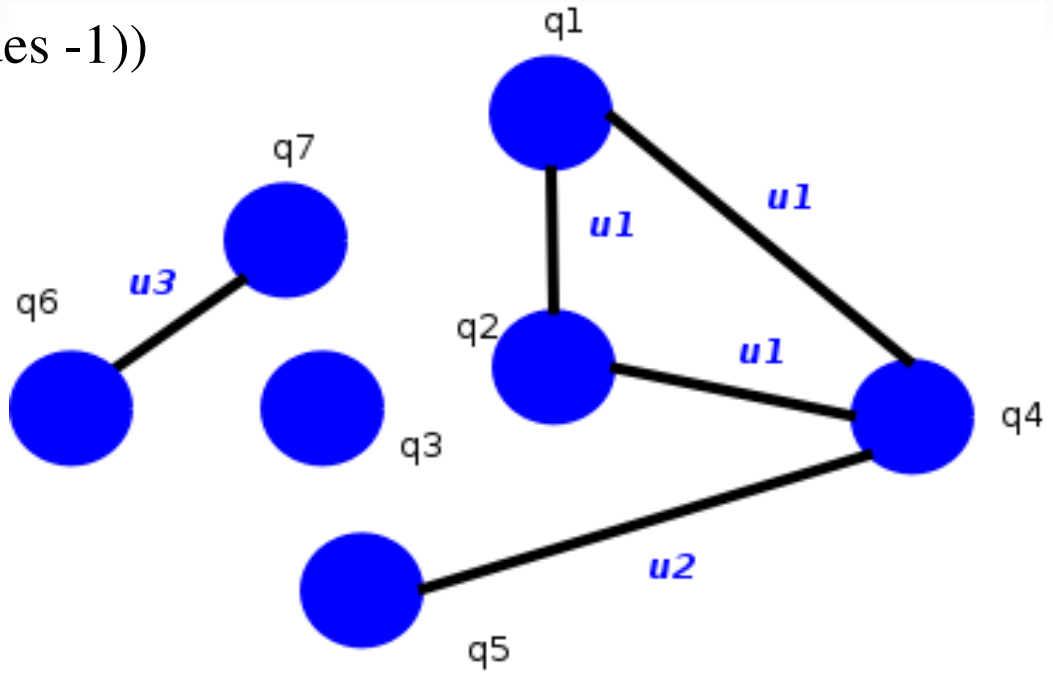
- Prevention of the disclosure of any information about a user (separate problem).

Outline

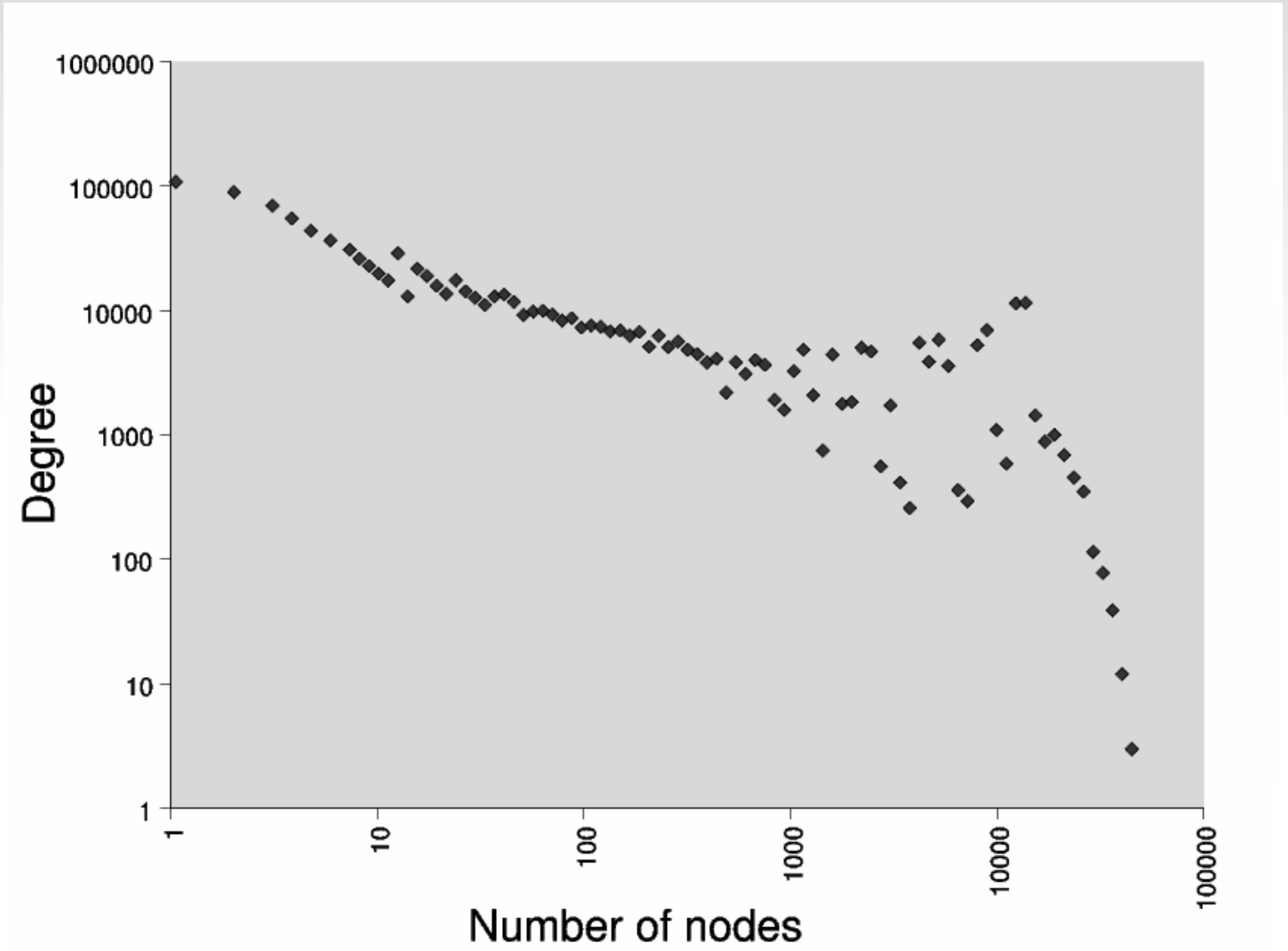
- ✓ We introduce a new privacy concern: *website privacy*.
- ✓ We show attacks that disclose confidential information about a website,
- ✓ and ways to prevent them.
- ✓ We propose a graph-based method for log anonymization.
- ■ We show a first experimental analysis over real data to evaluate one of our heuristics.

Evaluation

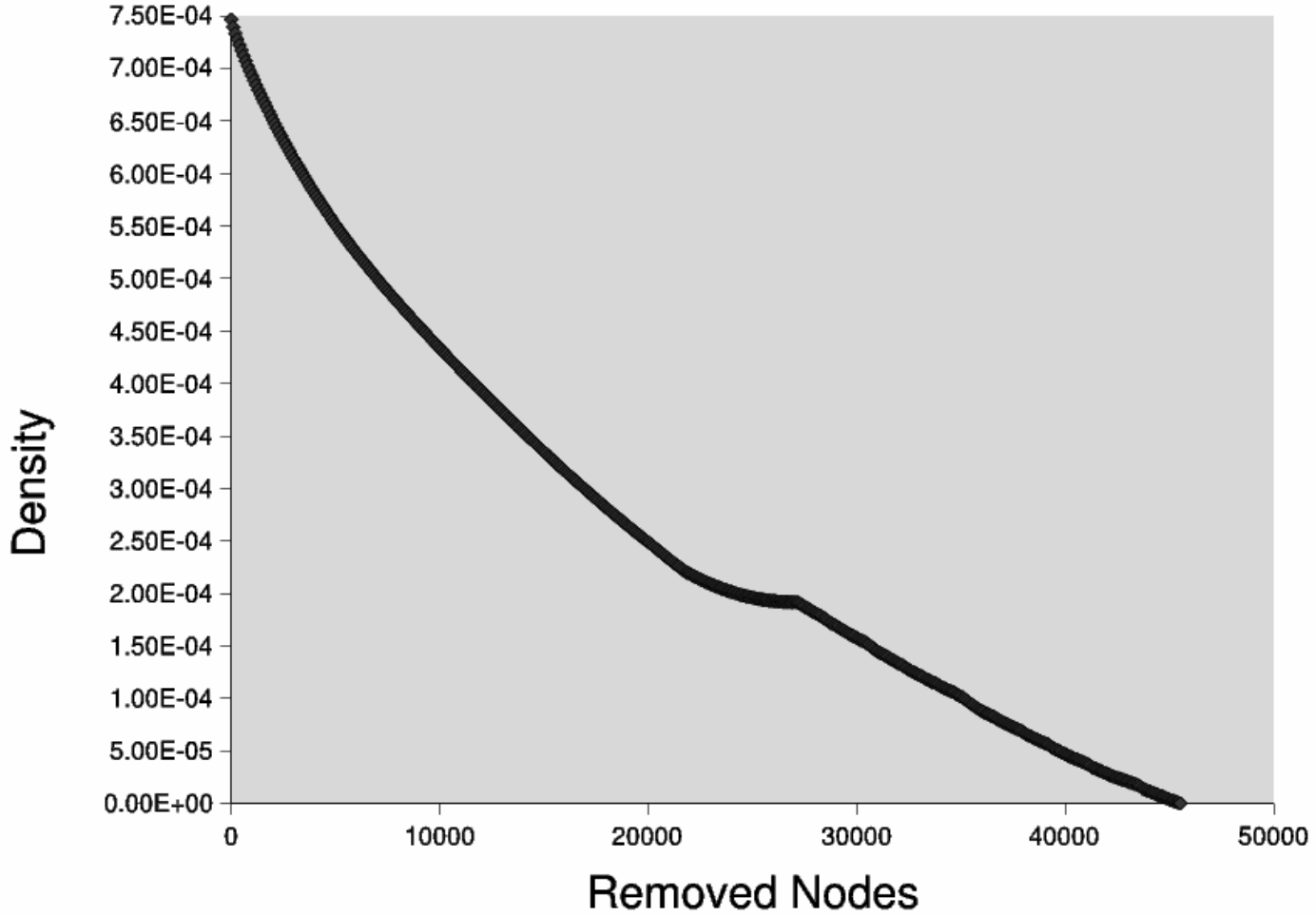
- Graph representation [R. Baeza-Yates & A. Tiberi, KDD'07], this can be computed fast.
- Yahoo! Log sample from 2005, with 3 million queries (nodes).
- Very disconnected to begin.
- *Density:* How likely it is to find an edge between any two nodes.
- *Density* = 2(# edges)/ (# nodes (# nodes -1))
- *Density* = 0.000089



Degree Distribution
(not a Power Law)



Density Vs. Number of Nodes Removed



Conclusions

- Our first results show that the anonymization of the query log against attack 3 required the removal of only a small fraction of the log:
  - 9.5% of the queries
  - 9.2% of the clicked documents
- The removed queries are infrequent, so the loss of information useful for mining is kept low.
- Mining tasks that do not require knowledge about a specific website can still be performed.
- Query logs usually show stable distributions, so our results are likely to be generalizable.
- This does not guarantee that the log is safe from other types of attacks.
- Full query log anonymization is an extremely difficult problem, there are no indications this will be achieved in the near future.
- However it is important to study website privacy preservation for applications that generate their results based on log data. To prevent adversarial attacks (e.g.: AdWords).