

Noise Injection for Search Privacy Protection

Shaozhi Ye

sye@ucdavis.edu

Department of Computer Science
University of California, Davis

Aug. 28, 2009

Joint work with Felix Wu, Raju Pandey, and Hao Chen.

Outline

- 1 Search Privacy
- 2 Noise Injection Model
- 3 Perfect Privacy Protection
- 4 Limited and Independent Noise
- 5 Future Work
- 6 Summary

1.1 Motivation

Threats to search users

- Large number of data mining algorithms + machines.
- Data retention window ranges from months to years.
- Vulnerable data sanitization designs and improper implementations:
 - AOL Gate: 20M queries from 650K "anonymized" users.
- Insider attack.

1.2 Search User Profiling

User identification

- IP address
- HTTP cookies
- Client-side tool: toolbar, desktop

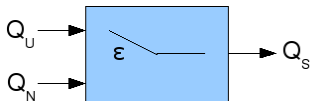
User profiling

- Queries
- Click-through
- Search preference: languages, categories
- Rich client side: toolbar, desktop

2.1 Noise injection Model

Noise Injection

- With probability ϵ , the user sends a true query Q_U
- With probability $1 - \epsilon$, the user sends a noise query Q_N



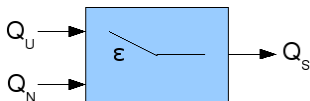
The search engine observes Q_S

$$\forall i \quad P(Q_S = q_i) = \epsilon P(Q_U = q_i) + (1 - \epsilon) P(Q_N = q_i)$$

2.2 Measure Privacy Breaches

Privacy breach

- The distribution of $Q_U \rightarrow$ user profiles.
- Mutual information $I(Q_S; Q_U)$



Problem

Find a Q_N such that $I(Q_S; Q_U)$ is minimized.

3. Perfect Privacy Protection

Theorem

$I(Q_S; Q_U) = 0$ *only if* $\epsilon \leq 1/N_Q$.

Corollary

Lower bound noise for a perfect protection:

$$E(|Q_n|) = \frac{1-\epsilon}{\epsilon} |Q_u| \geq (N_Q - 1) |Q_u|$$

Limitations

- Expensive:
 - Send the whole dictionary with each query.
- Limited bandwidth
- Search engines block users to prevent DoS attacks.
- Response delay:
 - Expected waiting time for each real query is $1/\epsilon$

4. Limited and Independent Noise

Let Q_u and Q_n be independent.

Optimization Problem

$$\arg \min_{\mathbf{n}} I(\mathbf{n}) \quad \text{w.r.t.} \quad \sum_i n_i = 1, \quad \forall i \quad n_i \geq 0$$

where $\mathbf{n} = (n_1, n_2, \dots, n_{N_Q})$.

Solution

- We prove I is a convex function of \mathbf{n} .
- Use Lagrange multipliers to solve the optimization problem.

4.2 A Special Case: $E(|Q_n|) = |Q_U|$

- Use Taylor series to replace the logarithm functions for an approximate solution.
- How close our solution is?
 - The objective function is convex.
 - Increasing the order of the Taylor series gets better accuracy.
- Caveat: Computational cost when N_Q is large.

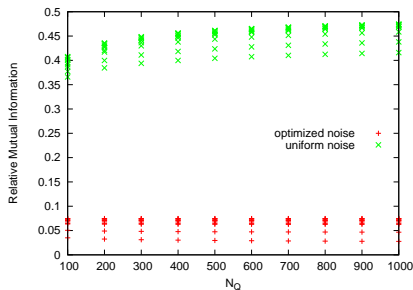
4.3 Simulation results

How to evaluate?

- The larger $H(Q_U)$ is, the larger $I(Q_S; Q_U)$ will be.
- Relative mutual information: $\frac{I(Q_S; Q_U)}{H(Q_U)}$.

Q_U : Power law distribution

The number of the i th most popular queries is proportional to $i^{-\alpha}$, $\alpha \in [1.0, 5.5]$.

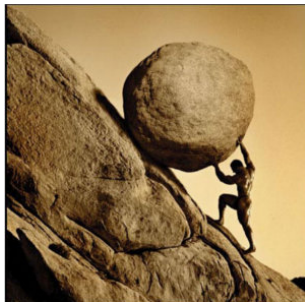


4.4 Applicability

- Privacy information is restricted within a relatively small sets of queries.
- Scalability
 - When N_Q increases, the protection of random noise gets worse while our solution does not exhibit such trend.
- Combining network based solutions with noise injection will help.

5. Future work

- Allow non-sensitive inferences.
- Allow attackers with external knowledge.
- Allow no prior knowledge on Q_U \rightarrow Adaptive noise generator.
- Have computational constraints for the attacker.



- Developed a noise injection model for search privacy protection.
- Proved the lower bound for the amount of noise queries required by a perfect privacy protection.
- Provided the optimal protection when noise is limited and independent of user queries.
- Computed an approximate solution for the case where same amount of noise is injected and evaluated our result with simulations.

Thanks!