Leveraging Query Sensitivity for Practical Private Web Search

Antoine Boutet, Albin Petit, Sonia Ben Mokhtar, Léa Laporte

▶ **To cite this version:**

Antoine Boutet, Albin Petit, Sonia Ben Mokhtar, Léa Laporte. Leveraging Query Sensitivity for Practical Private Web Search. Middleware, Dec 2016, Trento, Italy. 10.1145/3007592.3007595 . hal-01381995

HAL Id: hal-01381995
https://hal.science/hal-01381995v1

Submitted on 20 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Poster Abstract: Leveraging Query Sensitivity for Practical Private Web Search

Antoine Boutet, Albin Petit, Sonia Ben Mokhtar, Léa Laporte
Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
{antoine.boutet, albin.petit, sonia.benmokhtar, lea.laporte}@insa-lyon.fr

ABSTRACT

Several private Web search solutions have been proposed to preserve the user privacy while querying search engines. However, most of these solutions are costly in term of processing, network overhead and latency as they mostly rely on cryptographic techniques and/or the generation of fake requests. Furthermore, all these solutions protect all queries similarly, ignoring whether the original request contains sensitive content (e.g., religious, political or sexual orientation) or not. Based on an analysis of a real dataset of Web search requests, we show that queries related to sensitive matters are in practice a minority. As a consequence, protecting all queries similarly results in poor performance as a large number of queries get overprotected.

In this paper, we propose a request sensitivity assessment module that we use for improving the practicability of existing private web search solutions. We assess the sensitivity of a request in two phases: a semantic sensitivity analysis (based on the topic of the query) and a request linkability analysis (based on the similarity between the current query and the query history of the requester). Finally, the sensitivity assessment is used to adapt the level of protection of a given query according to its identified degree of sensitivity: the more sensitive a query is, the more protected it will be.

Experiments with a real dataset show that our approach can improve the performance of state-of-the-arts private Web search solutions by reducing the number of queries overprotected, while ensuring a similar level of privacy to the users, making them more likely to be used in practice.

CCS Concepts

•**Security and privacy** → **Privacy protections;**

1. INTRODUCTION

Search engines have become an essential service for finding content on the Internet. However, by regularly querying these services, users disclose a large amount of their personal data. To limit the disclosure of users' personal information, many private Web search solutions have been proposed. There are mainly two types of approaches, the solutions obfuscating the queries or obfuscating the identity of the requester. For instance, Tor [2] uses anonymous communication to hide the identity of the requester behind a chain of proxies, while GooPIR [3] generates k fake queries to hide among $(k+1)$ queries the real query of user. A recent study [5] show that each mechanism separately are not enough to efficiently protect users, and proposed to combine both mechanisms. However, this solution is costly in term of processing and network overheads. Indeed it relies on cryptographic operations and the generation of number of fake queries, which consequently induces loss of performance.

Furthermore, existing private Web search solutions protect all queries similarly regardless the sensitivity of the query. Based on a real dataset of query log and a crowdsourcing campaign to collect the human judgments regarding the semantically sensitive queries, we showed that overall sensitive queries are in practice a minority. The content oblivious nature of existing private Web search solutions can lead to an over-protection of non sensitive queries and an under-protection or sensitive queries. In this work, we explore an adaptive query protection driven by the sensitivity of the query. Specifically, we have identified two types of sensitive queries. The first one are the semantically sensitive queries. While there is no universally adopted definition of what constitutes sensitive content, the privacy policy of most service providers tend to define sensitive personal information with similar categories (e.g., religious beliefs, political orientation, sexual life, or medical facts). Here, we consider a score to quantify the risk that a query is semantically sensitive compared to categories pre-selected as sensitive by its associated requester. The second type of sensitive queries are those which can be linked to a user profile (i.e., sensitive to a de-anonymisation attack), named linkability sensitive queries. To quantify the risk that a query is linkability sensitive, we consider a score reflecting the similarity between the query and past queries of the associated user.

In this poster, we present a sensitivity assessment module which dynamically protects users' queries according to their level of sensitivity. This sensitivity assessment takes into account two sensitive matters, the semantic content and the linkability of the query. To protect a given query, the proposed module combines obfuscation (i.e., fake queries) and anonymous communication. Each of these mechanisms is controlled by the actual level of sensitivity of the query, the more sensitive it is, the more protected it will get. More precisely, the number and the generation of fake query depends

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Middleware Posters and Demos '16 December 12-16 2016, Trento, Italy

© 2016 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4666-5/16/12.

DOI: http://dx.doi.org/10.1145/3007592.3007595

on the semantic risk analysis, while the number of proxies part of the anonymous channel depends on the linkability risk analysis.

2. SENSITIVITY ASSESSMENT

Our sensitivity assessment module analyses the risk that a query is semantically and linkability sensitive.

2.1 Semantic Risk Analysis

The semantic risk analysis provides a score in $[0, 1]$ which determines how the query is related to the topics identified as sensitive by the user. To achieve that we use two libraries: (i) Wordnet, a lexical database, and (ii) eXtended WordNet Domains, a mapping of WordNet synsets to domain labels. First the query q is split into keywords. For each keyword, we collect all possible synsets (i.e., all possible meanings of one word in Wordnet) and return the most probable ones using a graph-based disambiguation method. Then, we map each synset to multiple categories using the eXtended WordNet Domains library. Finally, we use all these mappings to compute and return the probability that the query is related to the categories defined as sensitive by the requester.

2.2 Linkability Risk Analysis

The goal of the linkability risk analysis is to provide a score in $[0, 1]$ to determine if the query is sensitive to a de-anonymisation attack. In such an attack, an adversary tries to link an anonymous query to a specific user by measuring the distance between the query and background knowledge collected for each user (part of their query history in our case). This attack is successful if the query is close enough to past queries of the user. To quantify the linkability risk, we first represent the query q in a vector space model that has for values either 0 (the term is not in the query) or 1 (the term is in the query). We then compute and return the cosine similarity between the vector associated to the query and the vector representing past queries.

3. DYNAMIC QUERY PROTECTION

Our sensitivity assessment module dynamically adapts the query protection. More precisely, this assessment controls both the obfuscation and the anonymous communication.

Query obfuscation is achieved through the generation of k fake queries. To dynamically define the value of k for each query, we use a linear projection between the score returned by the semantic risk analysis in $[0, 1]$ and the number of fake queries in $[0, k_{max}]$. To dynamically define the number of proxies in the anonymous channel, named p, we use a linear projection between the score returned by the linkability risk analysis in $[0, 1]$ and p in $[1, p_{max}]$. Therefore, depending on the level of sensitivity of the query, the protection can be achieved through only one proxy (i.e., for non sensitive queries), or through multiple fake queries and a chain of multiple proxies (i.e., for the most sensitive queries).

4. PRELIMINARY RESULTS

We evaluate our sensitive assessment module using a dataset of real queries from AOL search. Our experiments show that the semantic assessment lead to the identification of semantically sensitive queries with a recall of 87.9% while the linkability assessment classifies linkable queries with a recall of 94.9%. Furthermore, we show that adapting our module to existing private Web search solution (i.e., PEAS) reduces by 40.3% the number of queries overprotected and consequently significantly decreases the latency and the bandwidth consumption. Lastly, we show that this performance improvement is achieved while maintaining a similar level of privacy (i.e., sensitive information captured by the search engine) and utility (i.e., quality of the results presented to users).

5. BACKGROUND ON SENSITIVITY

Privacy policies of online services tend to more and more take into consideration the sensitive nature of the information. For instance, the privacy policy of Quora (a question-and-answer service) offers to users the possibility to choose whether to answer questions anonymously or with their names attached. Specifically, this privacy feature has been recently analysed in [4] to develop a classifier to predict the sensitivity of individuals posts. Another recent work [1] provides a metric to identify the most susceptible users who are exposed to a personal sensitive state (e.g., pregnant or afflicted by depression) from their posts created in online communities. To identify sensitive topics, LDA models are build from large data collections specialized in the considered sensitive state. The similarity between the information published by users and the most salient words associated to sensitive topics is then performed to identify the more exposed users.

6. CONCLUSIONS

We propose a sensitivity assessment module to dynamically adapt the protection of each query according to its level of sensitivity. The protection is achieved by combining obfuscation through fake queries and anonymous communications. The number of fake queries is controlled by the sensitivity level of the semantic content of the query, and the number of relay in the proxy chain is controlled by the sensitivity level of linkability of the query to its requester. Preliminary results show that the proposed solution provides an important performance improvement while maintaining a similar privacy and utility to state-of-the-arts solutions.

7. REFERENCES

[1] J. A. Biega, K. P. Gummadi, I. Mele, D. Milchevski, C. Tryfonopoulos, and G. Weikum. R-susceptibility: An ir-centric approach to assessing privacy risks for users in online communities. In *SIGIR*, pages 365–374, 2016.

[2] R. Dingledine, N. Mathewson, and P. Syverson. Tor: The second-generation onion router. In *USENIX Security'04*, pages 21–21, 2004.

[3] J. Domingo-Ferrer, A. Solanas, and J. Castellà-Roca. h(k)-private information retrieval from privacy-uncooperative queryable databases. *Online Information Review*, 33(4):720–744, 2009.

[4] S. T. Peddinti, A. Korolova, E. Bursztein, and G. Sampemane. Cloak and swagger: Understanding data sensitivity through the lens of user anonymity. *S&P*, 2014.

[5] A. Petit, T. Cerqueus, S. Ben Mokhtar, L. Brunie, and H. Kosch. PEAS: Private, Efficient and Accurate Web Search. In *IEEE TrustCom'15*, Helsinki, Finland, Aug. 2015.