Université
franco-allemande
Deutsch-Französische
Hochschule

Introducing Privacy In Current Web Search Engines

by Albin Petit

Abstract

During the last few years, the technological progress in collecting, storing and processing a large quantity of data for a reasonable cost has raised serious privacy issues. Many online services (e.g., Facebook, Google) now have the ability to profile their customers to offer them targeted advertising. The revelation by Edward Snowden in May 2013 of a massive global surveillance program run by the NSA demonstrates that intelligence services are also collecting and exploiting a large quantity of personal data. Therefore, on a daily basis, many entities spy on users interactions, but most Internet users do not measure the extent of the collection. The main reason is that online services do not offer their users the possibility to access the collected data. In addition, their terms of service do not precisely inform users about what data is collected, and the purpose of the collection.

Privacy concerns many areas, but is especially important in frequently used websites like search engines (e.g., Google, Bing, Yahoo!). These services allow users to retrieve relevant content from an increasing amount of data published on the Internet. The good quality of their results comes from the exploitation of user personal data. As a direct consequence, search engines are likely to gather and store sensitive information about individual users (e.g., interests, political and religious orientations, health condition). In this context, developing solutions to enable users to query these search engines in a privacy-preserving way is becoming increasingly important.

In this thesis, we introduce SimAttack an attack against existing solutions to query a search engine in a privacy-preserving way. This attack aims at retrieving the original user query by exploiting unprotected user queries previously collected by an adversary. We use SimAttack to assess the robustness of three representative state-of-the-art privacy-preserving solutions. We show that these solutions are not satisfactory to protect the user privacy.

We therefore develop PEAS a new protection mechanism that better protects the user privacy (according to SimAttack). This solution leverages two types of protection: hiding the identity of the user and masking her queries. The former is achieved by ciphering and sending queries through a succession of two nodes, while the latter hides queries by combining them with several fake queries. The main challenge in our approach is to generate realistic fake queries. We solve it by generating queries that could have been sent by other users in the system.

Finally, we present mechanisms to identify sensitive queries. Our goal is to adapt existing protection mechanisms to protect sensitive queries only, and thus save resources (e.g., CPU, RAM). Indeed, a common query on a cake recipe does not need the same protection as a query on an HIV infection. We design two modules to identify sensitive queries and deploy them on real protection mechanisms. We establish empirically that adapting existing protection mechanisms dramatically improves their performance.