Demonstration of X-Search performances

Sandrine Juillard

Abstract—**Privacy on internet became a major issue these past years. The rapid enhancement of data analytics have lead to a new economic model base on the analyze of user behavior to personalize advertising.[1] X-Search is a proxy that aim to protect user from curious web search engine. The novelty of X-Search protocol [2] compared to other common secure web-search protocol is the use of Soft Guard Extention enclaves (SGX). As part of the "Parcours Ingénieur-Docteur", this report aim to implement a prototype base on X-Seach protocol and re-evaluate some aspect of the solution. In the present case, we will focus on the accurency of the algorithms and the impact of SGX on the responsiveness of the solution in order to answere the two following question: "How to improve SGX reponsivness" and "In what extend X-Search protocol is accurate ".**

Codes available here

Index Terms—**security, SGX, web search, privacy**

I. Introduction

APPLICATIONS we use daily, produces and accumulates numerical data. By collecting one's data, and comparing them to other users, analytic is able to create an accurate portrait of one's personality and thus, found the most accurate way to reach the user. These methods can be assimilated with manipulation and have been the cause of many scandals [1]. Hence, private Web search has been an active research area in the last decade in order to counterbalance the numerous threats open due to the oversharing of users' search queries to private companies. The purpose of X-search is to protect user from Web Search Engine such as Bing, google or yahoo to collect data from the content of their queries. Basically, it provide a proxy used between the user and the search engine. The solution proposed aim to overwhelmed the responsiveness / robustness compromise compare to others existing solutions like Thor[3]. Evaluations done in the original paper showed how X-search provide a good compromise; while being 3 times faster than Thor in average, it guarantee the integrity of the anonymity and is very robust to Re-identification attacks (these points will be developed further in section II). However, the protocol does not provide a perfectly accurate response to a query, the result is not the exact same answer as if the query was sent directly to the search engine. Hence the accuracy of the protocol have been measured in the original paper to proved it's reliability. However, several questions remains and it is possible to go further in the analyse of X-search performance. Two questions will be answered in this paper: the first one is "How X-search can be optimised" and the seconde one is "In what extend it is accurate". Hence, the first aspect to be evaluated will be the responsiveness. By knowing the time delay taken by each step of the protocol, it is possible to determine what step is the longest and should be optimised. Moreover, it is possible to judge, from the part of time taken by SGX, the reliability of using it in such context where responsiveness is essential. The second evaluation will focused on the accuracy of X-search, by evaluating it thought various dataset.

II. Background

A. Private Web Search

Private web search field is aiming to protect users against curious web search engine. On the contrary of cyber-security, the main treat is not hackers or to find potential security issues in a system, but to enable user to use web search services while preventing them to collect their data. Private Web Search existing solutions can be classified in three main categories.

Unlikability: In this category , privacy is guaranteed by the anonymity. Thus, origin of queries can't be associated with a user. The main protocol in this category is the Onion Routing protocol, RAC and the Dissent protocol [4][5][6]. Onion Routing guarantee anonymity by sending the query through multiple nodes. The nodes' path is selecting randomly and hidden using secure encrypted keys. The set of keys form the 'onion'. When the query is receive by a node, they remove "one layer" by deciphering the query revived and forward it to the next node. Despite the protocol protect effectively the anonymity of its user, it suffer from re-identification attacks. Indeed, the fact the query themselves can be used to break the unlikability. Re-identification attacks is able to link a profile with queries using the high correlations between queries from the same user. In addition, the main inconvenient of this protocol is its slowness. Thor is one of the most popular implementation of this protocol. An other issue of this protocol is that it can be biased by untruthful/malicious nodes. However, protocol such as RAC, which is an adaptation of the onion routing protocol, prevent malicious nodes to broke the protocol, nodes are grouped in a virtual ring. A node have to broadcast messages to all nodes' rings including it's predecessor, thus, if a predecessor does not received a message, it is informed that a node dropped a message.

Indistinguishability: This category of solution aim to spoil user profiles by sending dummy request to drown real request or by directly altered real request. However, this techniques are still vulnerable to re identification attacks. In fact, the fake queries generated are usually very distinct compared to real one. Thus, the correlation between true queries and fake queries are low. (Exemple of Indistinguihability for privacy using geo-localisation devices [7])

Alternative Search Engine: This last categories propose secure web search. Generally, these search Engines proposed a fully encrypted protocol which, despite being perfectly effective, is not feasible in terms of performance for an normal user. [8]

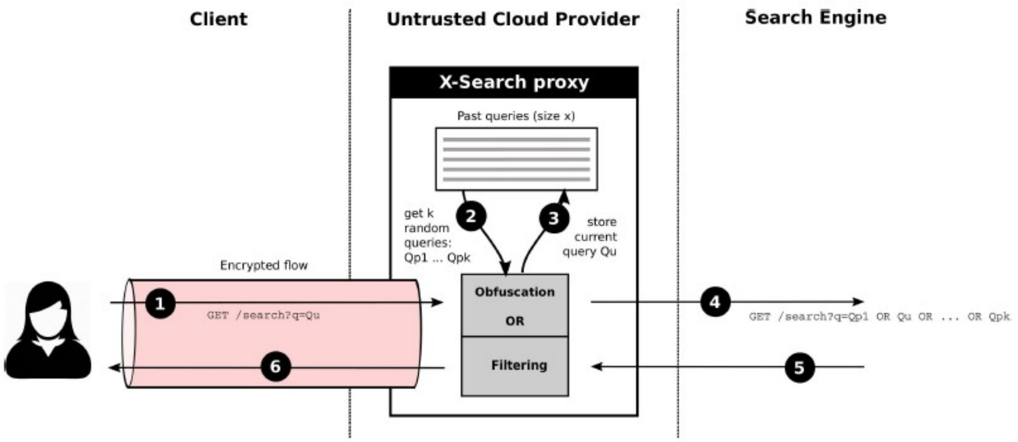

Fig. 1. The X-Search architecture and execution flow [2]

B. SGX Software

Software Guard Extensions (SGX) is a software dev kit developed with the 6generation of Intel processors, that enables the use of a secure execution environment named enclave. An Enclave is isolated from the device on witch it's executed by encrypting the totality of program data. Encryption keys are stored inside the processor itself, thus data can only be read inside the CPU. [9]. In this configuration, whatever the level of privilege accessed, data treated inside the enclave can't be obtained. Enclaves are an efficient protection for performing sensitive operation on an untrested platform. In practice, an untrusted application can create a trusted execution environment with SGX. The enclave initialised contain a set of functions. When a sensitive operation have to be made, the application will call a function in the enclave. The application and its enclave communicate trough messages named Ecall (Enclave Calls) and Ocall (Outside Calls). Ecalls is a call from the application to the enclave and Ocalls allow the enclave to call of function outside the enclave. Indeed, an enclaves can't do any operation that need a system call.[10]

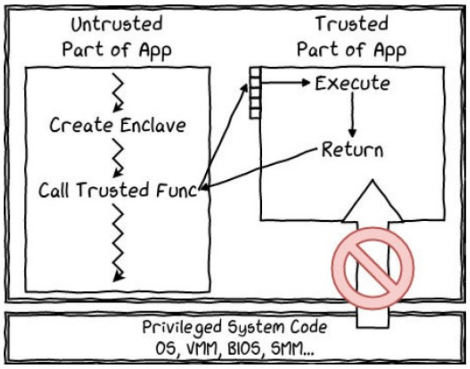

Fig. 2. Overview of an application using Soft Guard Extention [9]

C. Challenges in Private Web Search Area

```
Algorithm 2: Results filtering.
input: $Q_u$ : initial query,
       $pastQuery = \{Q_{p1}, \dots, Q_{pk}\}$ : set of past queries,
       $R$ : set of results for $Q_u \vee Q_{p1} \vee \cdots \vee Q_{pk}$.
$\bar{R} \leftarrow \emptyset$ ;
$q^+ \leftarrow \{Q_u, Q_{p1}, \dots, Q_{pk}\}$ ;
for $r \in R$ do
    for $q_i \in q^+$ do
        $score[q_i] \leftarrow$ nbCommonWords($q_i$, title($r$))
          + nbCommonWords($q_i$, desc($r$));
    if $score[Q_u] = \max_{q_i \in q^+} score[q_i]$ then
        $\bar{R} \leftarrow \bar{R} \cup \{r\}$ ;
return $\bar{R}$ ;
```

Fig. 3. Pseudo-Algorithm of the filtering Algorithm from the X-Search original paper

X-search propose a protocol combining unlikability and indistinguishably allowing good responsiveness and robustness to re-identification attacks. Indeed, thanks to Soft Guard Extension Enclave, the proxy can deploy on any cloud as a node that be will be guaranteed unbiased on any platforms regardless their trustfulness. And in addition of providing a proxy guaranteeing anonymity, it also protect user from re-identification attacks by drawing real request with real users request. In fact, past queries from user can be securely stored inside the enclave providing a dataset to assure indistinguishability of the query. Since both real and dummy queries are from real user, the correlation between users is impossible to determine. Hence, X-Search bring a swift and reliable solution to guaranteed privacy while using web search engine.

III. X-Search overview

The execution flow of X-search is depict figure 1:

$\boxed{1}$ The user send its request to a node hosting X-search proxy using an Encrypted flow. A specific broker is in charge of managing the encryption keys to enable a protected communication.

$\boxed{2}$ X-search proxy receive the query. From an history of past queries (witch update each time a request is made), the original queries of the user is combine with K randoms queries into a single one queries using the ”OR” operator. [11] This is the Obfuscation treatment.

3 The current user query is added anonymously to the database.

4 The obfuscated query is sent to a web search engine.

5 The result from a web query is a set of links, titles and description.The answer to an obfuscated query is the result from each queries mixed together. Results as selected by the number of words in common between the query and the titles and descriptions. The score is calculate for each queries that have been obfuscate. The result sent to the user is results that have more common word with the original query than any other queries 3

6 The result obtain (sets of links, titles and descriptions) is forward to the user. In addition, theses results are tampered by the proxy to remove any URL redirection to protect anonymity.

All X-search action on the proxy will take place inside the enclave. An ocall is required to reception the request and send it. The history of past queries have a maximum size, reliant on SGX protected memory capacity of 128 MB, which is equivalent to stored up to 1 million queries. Concerning the algorithms, it is important to mention that the filtering method does not provide constant number of results as outputs. Hence, asking an increased numbers of results maximise the probability of having a reliable number of outputs. In other words, the filter select the answer to the original query only, but it is possible that the first true results only appeared in the 10th position.

IV. Experimental Setup

A simplified implementation of X-search have been developed to evaluate specifically two aspects that will be detailed in section V. Some specificity of X-search have been deliberately ignore; the implementation is minimal to focus on evaluations requirement. The protocol implemented is depict in figure 4. The X-search protocol is deployed on an Azure virtual machine. The algorithms are identical to theses presented in the original paper. The main difference between the original protocol and this one is that the history of queries is static and it haven’t been deployed as a proxy bu rather as an app. Details of the implementation is describe Table I. This implementation will be compared with performance of two others variants of the implementation: A strictly identical implementation performed fully outside the enclave (referred as Algorithms implementation). And an implementation that do raw queries to the API (request + pacing).

A. Technical Problems encounter

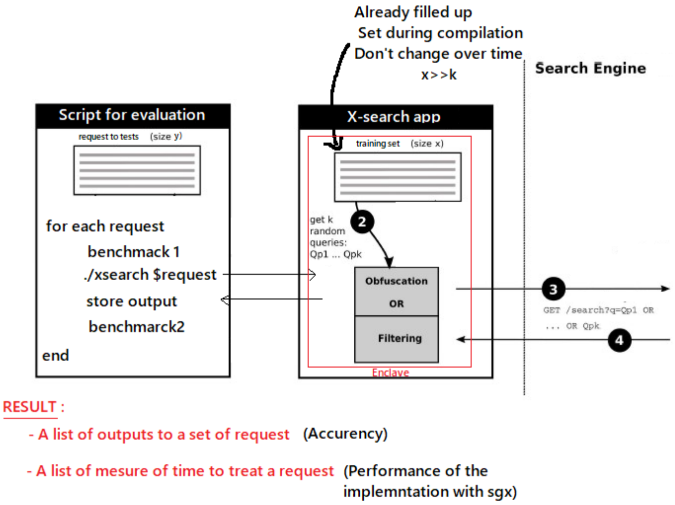

Fig. 4. New implementation’s overview

TABLE I
Summery table of implementation settings

	SGX Original	New implementation
Test dataset	Source: 100 most active users from AOL[14] Size: 2/3 of the database	Source: random queries from AOL Size: 5000 queries
History	Size: 1M queries Training: 2/3 of the database	Size:2500 Static / no training
Web seach engin	bing.com	Azure bing search API
Operating system	Ubuntu Desktop LTS 14.04	Ubuntu Server LTS 18.04
Processor	Intel® Core™ i7-6700	Intel® Xeon® Platinum 8272CL
RAM	8 GiB	3.5 GiB
Enclave	Regular mode	Simulation mode

Unfortunately, Virtual machine of azure enabling Soft guard extension are not available for trial subscription. The VM needed to run SGX is the ”Confidential compute” [12]. To get around this inconvenient evaluation will be run on a regular VM using the simulation mode of SGX but the time taken by the enclave will not be measured correctly as wanted. Indeed, the simulation mode works in the same way as the debug mode except the fact that true hardware is not exercised, instead the Intel SGX instructions are simulated in software [13]. Results from the original paper will be used to answer addressed questions, even if the implementation is not strictly identical. In addition, the Azure Bing API is mildly expensive. Hence, with a limited credit, the number of request by experiment had to be limited. (The cheaper subscription being approximately 3$ per 1000 queries).

V. Evaluation

In the original paper, the evaluation aim to determined if the objectives of responsiveness, robustness to re-identification attack are complete and if the accuracy is correct. Inspired by these previous evaluation, this section will present you the two new evaluations designed to analyse further X-Search.

A. Algorithm Accuracy

Measuring the accuracy of responses mean measuring if the responses given by the proxy X-search are equivalent to the responses given directly by the search engine. The accuracy of the proxy is measured with the same metrics used in the original paper which are define below.

$$precision = \frac{|R_{or} \cap R_{xs}|}{|R_{xs}|} \qquad recall = \frac{|R_{or} \cap R_{xs}|}{|R_{or}|}$$

The accuracy metrics, assesses the quality of the query results provided to users according to the results obtain by directly calling the web search engine. We consider the precision and the recall . Precision will evaluate the proportion of results from X-search that are actually related to the original query. It will be measured by the intersection between R_{or}, result from original query and R_{xs},result from X-search, divided by the number of queries given by X-search (i.e., correctness). The recall will measured the proportion of results from original results that are effectively transmitted to the user (i.e., completeness). R_{os} size i set to 10. The number of results taken as input of the filtering algorithm is set to 50. No information about these parameters have been found in the original paper.

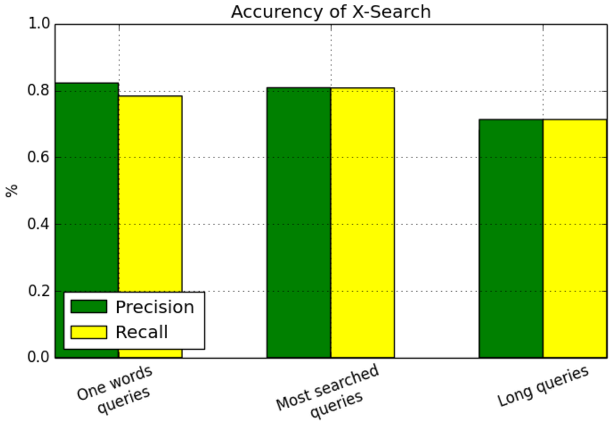

Fig. 5. Accuracy of X-Search by choices of dataset

In the original paper, queries have been selected from the 100 most active user and tested for different values of K. In order to see in what cases X-search protocol is weak and in which it is accurate, we tested different dataset classed by categories and test the precision and recall for each one. Categories are: One word queries, more than one word queries, 1000 most searched queries. Surprisingly, the percentage of recall and precision obtained is very different from the original paper. For K=2 (which mean that the original query is obfuscated with 1 query), the precision and recall are lower than 50%. These result can be raised by increasing the number of results taken as input of the filtering algorithm is set to 50 and lowering the size of R_{os}.

Furthermore, the experiment II put in light weaknesses of OR operator which can explain such low result. When a query contain mistake Bing automatically give results of the corrected with the mention ”Showing results for patti smith, Search instead for petti smith”. This mechanism does not works with obfuscated queries. Moreover, it is clear that ”google” is way more represented than ”Patti Smith” since it is a more common query. Hence, in order to works properly,X-search query have to be more common than dummy queries and should not contain mistakes and preferred short queries to long queries with several words.

TABLE II
STATISTICS OVER THE 10 FIRST RESULTS FOR DIFFERENT QUERIES

		Results for Patti smith	Results for google
No spelling mistake	*Patti Smith OR google*	30%	70%
One spelling mistake	*Petti Smith OR google*	0	100%
	Patti smith OR goole	90%	10%
Two spelling mistakes	*Petti smith OR goole*	10%	90%

B. System perofrmance with SGX

This evaluation 6 will focus on the responsiveness of the application. This mean measuring the time taken by X-search to answer a query. In the original paper, the time is measured for an end-user. In our implementation, no encrypted communication nor user friendly UX have been developed. We will focus on the delta of time added by SGX and the algorithms compare to a raw request to Bing using the Azure API [15]. Parcing of the responses from the web serach engine have been done in both implementation. This evaluation will provide an approximation of the impact of the usage of soft guard extension on the responsiveness. Hence we will be able to answer the question: ”What part of the X-Search protocol does take the most time” in order to better understand what part of X-search could be optimised to enhance responsiveness, the CDF (Cumulative Distributed Function) of time latency from several variant of implementation have been computed. As showed on figure 6, the use of SGX in simulation mode have a negligible impact on responsiveness compare to the algorithms. Indeed, the delay between the two implementation in average Δ_{sgx} is inferior to 0.01s. Hence, approximately 95% of $\Delta_{X-search}$ is caused by the treatment time of the algorithms Δ_{algo} (Filtering/Obfuscate).

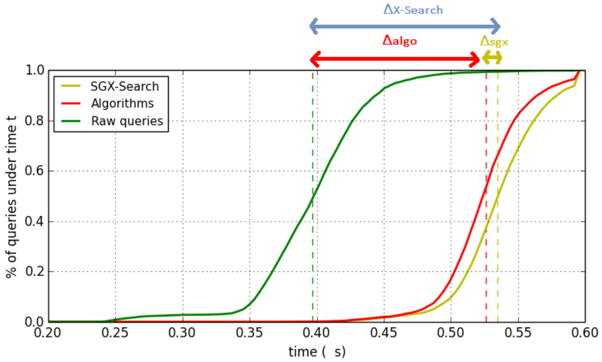

Fig. 6. CFD of the time taken by different of web search’s implementations

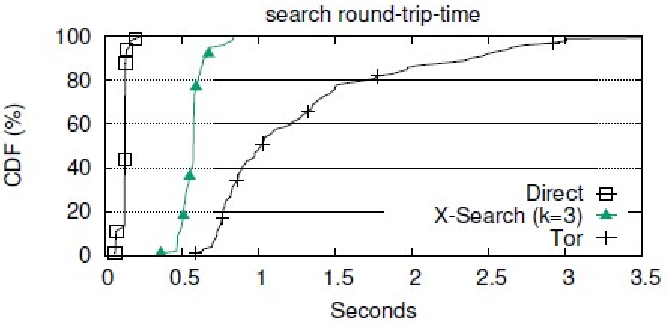

Fig. 7. Evaluation of X-Search responsiveness from the original paper

The result from the original paper showed close result from ours, around 0.5s. Hence, despite the two implementation are not strictly identical, it is possible to interpret SGX in simulation mode and in regular mode have close performances. Hence, part of SGX seems to be negletable compare to algorithms treatment.

VI. Conclusion

In the light of the experiments, SGX responsiveness performance reliability have been proved in such context as Web search, where speediness is a keen interest. Others SGX application support our conclusion. For instance, SGX have been used to encrypt entries of keyboard and mouse which require a treatment below 200ms [16]. Our experiment does not provide the part taken by the enclave since it was tested in simulation mode, but in addition of the evaluation from the original paper and the knowledge from article related to SGX, we can certainly assure that the time added by SGX is not a problem. However, performance of an application can be reduced by numerous ecall and ocall. In fact, such operation perform context switches and solution to optimised this aspect have already been examined. Among others proposition, [17] suggest an alternative method of communication between the enclave and the untrusted part of the application, using shared memory. Furthermore, regarding X-search performance on itself, the major part of the delay discrepancy seems to not coming from the algorithm neither SGX but from the parsing of raw requests performed. Concerning the accuracy of the solution, result are reliable provided that queries are short, common and does not contain spelling mistakes. Indeed, the use of the ”OR” operator provide less significant responses [18]. Despite this X-search seems to be impractical, the ideas of using Soft Guard extension for private Web search seems to be interesting since it provide the guarantee that nodes aren’t malicious. The idea have been developed to adapt Tor browser [19] and to enhance crytpo-currency protocols [20]. Hence as said in the original paper, if with time every personal PC became compatible with SGX, such protocol using SGX to guaranteed trustfulness of nodes could be democratise.

References

[1] S.Juillard, “Why are we scared of big data,” *Enseirb Matemca - EC9LC301 LV1 Anglais*, 2020.

[2] A. Felber M.Pasin R.Pires V.Schiavoni S.Ben Mokhtar, “X-search : Revisitingprivateweb search using intel sgx,” *Association for Computing Machinery*, 2017.

[3] B. E. J. H. E. K. J. S. M. den Toom R.M. van der Laan, “ Thirteen Years of Tor Attacks,” *Computer Science, Delft University of Technology*, 2018.

[4] P. S. Roger Dingledine, Nick Mathewson, “Tor: The Second-Generation Onion Router,” *Conference on Distributed Computing Systems, USENIX Security Symposium*, 2004.

[5] S. B. M. G. B. A. D. V. Q. A. Shoker, “RAC: a Freerider-resilient, Scalable, Anonymous Communication Protocol,” *Conference on Distributed Computing Systems*, 2013.

[6] V. S. Bryan Ford, Joan Feigenbaum, “Dissent accountable anonymous group communication,” *Yale University*, 2011.

[7] M. A. N. B. K. C. C. Palamidessi, “Geo-Indistinguishability: Differential Privacy for Location-Based Systems,” 2013.

[8] C. Gentry, “A fully homomorphic encryption scheme ,” September 2009.

[9] A. Adamski, “Overview of Intel SGX.Quarkslab’s blog,” 2018.

[10] S.Juillard, “Parcours ingénieur-docteur,” *Enseirb Matemca*, 2020.

[11] B. documentation, “Advanced search options,” 2020.

[12] A. Documentation, “Quickstart: Deploy an Azure confidential computing VM in the Azure portal,” 2020.

[13] Intel, “How to run intel® software guard extensions’ simulation mode,” 2016.

[14] C. T. G. Pass, A. Chowdhury, “A Picture of Search,” *he First International Conference on Scalable Information Systems*, 2006.

[15] A. documentation, “ What is the Bing Web Search API,” 2020.

[16] C. H. C. for Information Security, “Fidelius: Protecting user secrets from compromised browsers,” 2019.

[17] T. A. Ofir Weisse, Valeria Bertacco, “Regaining lost cycles with hotcalls: A fast interface for sgx secure enclaves,” 2017.

[18] B. J. J. Caroline M. Eastman, “The Appropriate (and Inappropriate) Use of Query Operators and Their Effect on Web Search Results ,” *ASIS&T*, 2004.

[19] S. Kim, J. Han, J. Ha, T. Kim, and D. Han, “Sgx-tor: A secure and practical tor anonymity network with sgx enclaves,” *IEEE/ACM Transactions on Networking*, vol. 26, no. 5, pp. 2174–2187, 2018.

[20] I. M. K. W. M. Schneider and K. K. E. Z. G. Karame, “Bite: Bitcoin lightweight client privacy using trusted execution,” *28th USENIX Security Symposium*, 2019.