International Journal of Communication Networks and Security

Volume 2 | Issue 2 Article 8

April 2013

PUBLISHING SEARCH LOGS PRIVACY GUARANTEE FOR USER SENSITIVE INFORMATION

J.ARUNA SANTHI
Mahatma Gandhi Institute of Technology, Gandipet, Hyderabad, Shanthi.it04@gmail.com

CH.LAKSHMI KUMARI
Mahatma Gandhi Institute of Technology, Gandipet, Hyderabad., lakshmi.itsmine@gmail.com

NANDITHA B
Mahatma Gandhi Institute of Technology, Gandipet, Hyderabad., nanditha.boddu@gmail.com

B .MEENAKSHI
Mahatma Gandhi Institute of Technology, Gandipet, Hyderabad, india3005@gmail.com

Recommended Citation
SANTHI, J.ARUNA; KUMARI, CH.LAKSHMI; B, NANDITHA; and .MEENAKSHI, B (2013) "PUBLISHING SEARCH LOGS PRIVACY GUARANTEE FOR USER SENSITIVE INFORMATION," *International Journal of Communication Networks and Security*: Vol. 2 : Iss. 2 , Article 8.
DOI: 10.47893/IJCNS.2013.1082
Available at: https://www.interscience.in/ijcns/vol2/iss2/8

PUBLISHING SEARCH LOGS PRIVACY GUARANTEE FOR USER SENSITIVE INFORMATION

J.ARUNA SANTHI[1], CH.LAKSHMI KUMARI[2], NANDITHA[3], B.MEENAKSHI[4]

[1,2,3,4]Mahatma Gandhi Institute of Technology, Gandipet, Hyderabad.
E-mail: Shanthi.it04@gmail.com, lakshmi.itsmine@gmail.com,nanditha.boddu@gmail.com,india3005@gmail.com

Abstract- Search Engine companies maintain the search log to store the histories of their users search queries. These search logs are gold mines for researchers. However, Search engine companies take care of publishing search log in order to provide privacy for user's sensitive information. In this paper we analyze algorithm for publishing frequent keywords, Queries, and Clicks of a search log. Before Zealous algorithm, we discuss how different variants of anonymity failed to provide good utility (publishing frequent items) and strong privacy for the search logs. And also this paper includes how zealous algorithm provides good utility and strong privacy for publishing search logs.

Keywords*- Search log, Information storage and retrieval, security, privacy, publishing frequent items.*

1. INTRODUCTION

Search engines play a crucial role in the navigation through the vastness of the web.Today's search engines do mine information about their users. They store the queries , clicks , IP-addresses , and other information about the interactions with users in what is called a search log. Search logs contain valuable information that search engines use to tailor their services better to their users' needs . They enable the discovery of trends , patterns in the search behavior of users, and they can be used in the development and testing of new algorithms to improve search performance and quality. Scientists all around the world would like to tap this gold mine for their own research ; search engine companies, however do not release them because they contain sensitive information about their users , for example searches for diseases, lifestyle choices, personal tastes, and political affiliations .

The only release of a search log happened in 2006 by AOL, and it went into the annals of tech history as one of the great debacles in the search industry. AOL published three months of search logs of 650,000 users. The only measure to protect user privacy was the replacement of user-ids with random numbers-utterly insufficient protection as the New York Times showed by identifying a user from Lilburn, Georgia [4], whose search queries not only contained identifying information but also sensitive information about her friend's ailments.

2. PRELIMINARIES

2.1 Search logs
Search engines such as Bing, Google, or yahoo log interactions with their users. When a user submits a query and clicks on one or more results, a new entry is added to the search log. Without loss of generality, we assume that a search log has the following schema {USER-ID, QUERY. TIME, CLICKS}, where a USER-ID identifies a user, a QUERY is a set of keywords, and CLICKS is a list of urls that the user clicked on. The user-id can be determined in various ways; for example, through cookies, IP addresses, or user accounts. A user history or search history consists of all search entries from a user. Such a history is usually partitioned into sessions containing similar queries; A query pair consists of two subsequent queries from the same user within the same session.

We say that a user history contains a keyword k if there exists a search log entry such that k is a keyword in the query of the search log. A keyword histogram of a search log S records for each keyword k the number of users c_k whose search history in S contains k. A keyword histogram is thus a set of pairs (k, c_k). We define the query histogram, the query pair histogram, and the click histogram similarly. We classify a keyword, query, consecutive query and click in a histogram to be frequent if its count exceeds some pre-define threshold T.

2.2 Disclosure Limitation methods for publishing search logs

1. K – anonymity
2. ϵ-Differential privacy
3. Probabilistic differential privacy

2.2.1 K- Anonymity: There are different variances of k- anonymity for search logs. Adar proposes to partition search log in to sessions and then to discard queries that are associated with fewer than k-different user-ids .In each session, the user-id is then replaced by a random number. Output of Adders algorithm called as K-query anonymous search log.Nabir add or delete keywords from session until each session contains same keywords in the search log, following by a replacement of the user-id by random number.

Output of this algorithm is called as k-session anonymous search log.

Naughton generalize keywords by taking their prefix until each keyword is part of at least k-search histories and publish a histogram of the partially generalized keywords. The output of algorithm is called k-keyword anonymous search log. All the variants of k-anonymities are insufficient in the light of attackers who can actively influence the search log.

2.2.2 Insufficiency of Anonymity:

K- Anonymity and its variants prevent an attacker from uniquely identifying the user that corresponds to a search history in the sanitized search log. However, even without unique identification of user, an attacker can gather the keywords or queries used by the user. K- Anonymity does not protect against this severe information disclosure.

2.2.3 Impossibility of Differential privacy:

Differential privacy provides much stronger privacy than k-utility, however

It is impossible to achieve good utility(publishing frequent items).

2.2.4. Probabilistic Differential Privacy

Probabilistic Differential Privacy achieves epsilon Differential Privacy with high probability.However, under realistic settings, no differentially private algorithm produce a sanitized search log with reasonable utility.In other words no differentially private algorithm can be accurate for both very frequent and very infrequent query-pairs.

3. ACHIEVING PRIVACY

3.1 Publishing frequent items of search log using Zealous algorithm :

Korolova , Michaela Gotz et al. developed Zealous algorithm . Zealous algorithm ensures strong privacy with good utility by following two phases of frame work. In the first phase, Zealous generates a histogram of items in the input search log and then removes from the histogram the items with frequencies below a threshold. In the second phase, Zealous adds noise to the histogram counts, and eliminates the items whose noisy frequencies are smaller than another threshold. The resulting histogram (sanitized histogram) is returned as the output of Zealous algorithm.

3.2 Zealous Algorithm for publishing Frequent Items of a search Log

Input:

Search log S positive numbers m , T, T′

Where m =Maximum no. of items from user.

T=First Threshold value on original histogram.

T′=Second Threshold Value .

Step1: For each user u select a set s_u of up to m distinct items from search history.

Step2: Based on the selected items, create a histogram consisting of pairs(k,c_k) where k denotes an item (keyword or query) and c_k denotes the number of users u that have k in their search history s_u. We call this histogram as Original histogram.

Step3: Delete from the histogram the pairs (k, c_k) with count c_k smaller than T(Threshold value).

Step4: For each pair (k , c_k) in the histogram, sample a random number n_k and add n_k to the count c_k.Resulting in a noisy count $c_K' \leftarrow c_k + n_k$.

Step5: Delete from the histogram the pairs (k, c_k) with noisy counts $c_K' \leq T'$.

Step6: Publish the remaining items and their noisy counts (Sanitized histogram).

4. RESULT AND DISCUSSION

Table1: Data set

User-id	Keyword	Link	Date Time
jstein	java	www.roseindia.com	2012/11/12:10:14
Carlu	C++	www.cpp.com	2012/11/12:10:14
Bob	java	www.roseindia.com	2012/11/12:10:14
Smith	mobile	www.ebay.com	2012/11/12:10:14
Smith	mobile	www.ebay.com	2012/12/12 10:12
Jestin	laptop	www.ebay.com	2012/11/12 10:17
alice	dwdm	www.wikipedia.org	2012/11/12/5:00
carlu	laptop	www.Tradus.com	2012/11/12 6:12

For the Table1 dataset, we generated base histogram, which consists the number of clicks per keyword i.e. the keyword searched by users (including duplicate users).

4.1 BASE HISTOGRAM

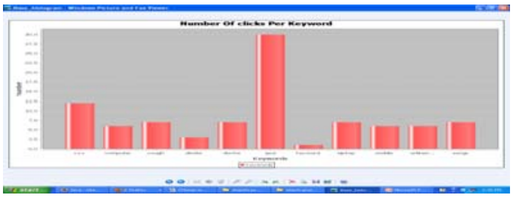

Fig.1

As a second step(step:2) of zealous we create a original histogram which generates the number of users per keyword(without duplicates),and it will generates corresponding log file which in the .csv format.From the algorithm.

4.2 ORIGINAL HISTOGRAM

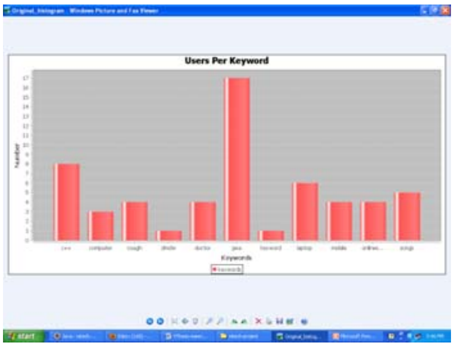

Fig.2

As a third step of zealous,we will delete the pairs from histogram which are less than the pre defined threshold value.here we have choosen the threshold value T=3.

4.3 Original Histogram (after applying T)

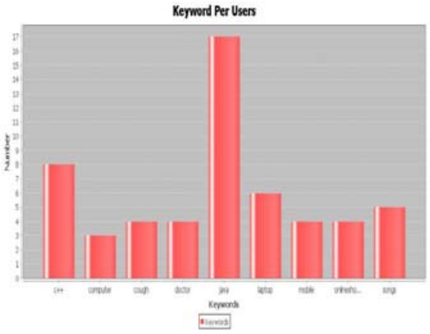

Fig.3

After adding noisy to original histogram looks like this.And we added randomly some noisy users for the search log of original histogram.

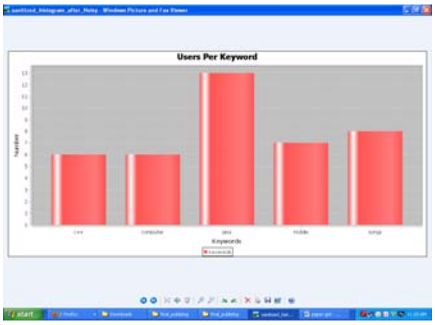

Fig.4

Sanitized histogram after removing noisy counts by applying T′ .

Choosing a value of T′ depends on the noisy count how much we have added.

4.4 Sanitized Histogram

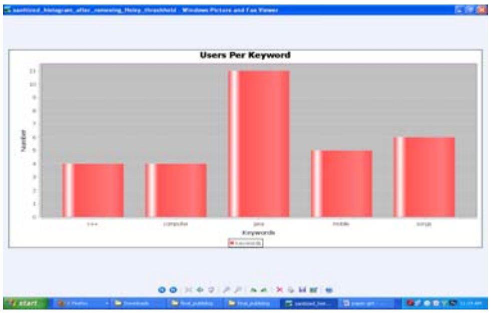

Fig.5

Finally after removing noisy counts our final search log contains records as like original histogram,but this final publishing search log vary with original histogram as it contains the noisy users with same frequencies.

4.5 Publishing sanitized histogram

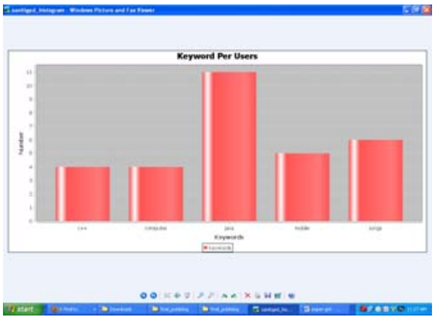

Fig.6

5. CONCLUSION

We analyzed algorithm for publishing frequent keywords, and Clicks of a search log. we discussed how different variants of anonymity failed to provide good utility and strong privacy for the search logs. This paper concludes with how zealous algorithm provides good utility and strong privacy for publishing search logs. And also we have chosen different values for parameters m , T, T′ to show different variations of utility (publishing frequent items).

6. FUTURE ENHANCEMENTS

We evaluate the performance algorithm in two ways. First, we measure how well the output of the algorithms preserves selected statistics of the original search log. Second we can consider two real applications from information retrieval community to evaluate the utility of Zealous. Index caching as a representative application for search performance, and query substitution as a representative application for search quality. Evaluating the output of Zealous algorithm with these two applications will help us to fully understand the performance of Zealous in an application context.

REFERENCES

[1] E. Adar, "User 4xxxxx9: Anonymizing Query Logs," Proc. World Wide Web (WWW) Workshop Query Log Analysis, 2007.

[2] R. Baez-Yates, "Web Usage Mining in Search Engines," Web Mining: Applications and Techniques, Idea Group, 2004.

[3] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar, "Privacy Accuracy and Consistency Too: A Holistic Solution to Contingency Table Release," Proc. ACM IGMODSIGACT- SIGART Symp. Principles of Database Systems (PODS), 2007.

[4] M. Barbaro and T. Zeller, "A Face is Exposed for AOL Searcher No. 4417749," New York Times, http://www.nytimes.com/ 2006/08/09/technology/09aol.html?

[5] A. Blum, K. Ligett, and A. Roth, "A Learning Theory Approach to Non-Interactive Database Privacy," Proc. 40th Ann. ACM Symp. Theory of Computing (STOC), pp. 609-618, 2008.

[6] J. Brickell and V. Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2008.

◈◈◈