Volume 2, Issue 8, August 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Comparative Study on Privacy by Search Engines while Publishing Search Logs

[1]Nayeneni Lavanya. [2]Mrs. E.Sandhya Rani , 3Dr.P.Raja Prakash Rao
[1] M.TECH STUDENT OF TRR ENGINEERING COLLEGE, HYDERABAD, INDIA
[2] ASSISTANT PROFESSOR OF TRR ENGINEERING COLLEGE, HYDERABAD, INDIA
[3] PROFESSOR OF TRR ENGINEERING COLLEGE, HYDERABAD, INDIA

Abstract: ***Search engine companies collect the "database of intentions", the histories of their users' search queries. These search logs are a gold mine for researchers. Search engine companies, however, are wary of publishing search logs in order not to disclose sensitive information. In this paper we analyze algorithms for publishing frequent keywords, queries and clicks of a search log. We first show how methods that achieve variants of k-anonymity are vulnerable to active attacks. We then demonstrate that the stronger guarantee ensured by Є-differential privacy unfortunately does not provide any utility for this problem. We then propose a novel algorithm ZEALOUS and show how to set its parameters to achieve (Є, δ)-probabilistic privacy.***

Our paper concludes with a large experimental study using real applications where we compare ZEALOUS and previous work that achieves k-anonymity in search log publishing. Our results show that ZEALOUS yields comparable utility to k−anonymity while at the same time achieving much stronger privacy guarantees.

Keywords:ZEALOUS,searchlogs,tok−anonymity,**AynRand**

I. INTRODUCTION

Civilization is the progress toward a society of privacy. The savage's whole existence is public, ruled by the laws of his tribe. Civilization is the process of setting man free from men. —AynRand.

My favorite thing about the Internet is that you get to go into the private world of real creeps without having to smell them. —PennJillette.

Search engines play a crucial role in the navigation through the vastness of the web. Today's search engines do not just collect and index web pages, they also collect and 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 and anomalies 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 2007 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 friends' ailments.

The AOL search log release shows that simply replacing user–ids with random numbers does not prevent information disclosure. Other ad–hoc methods have been studied and found to be similarly insufficient, such as the removal of names, age, zip codes and other identifiers [14] and the replacement of keywords in search queries by random numbers [18].We compare formal methods of limiting disclosure when publishing frequent keywords, queries, and clicks of a search log. These methods we study vary in the guarantee of disclosure limitations they provide and in the amount of useful information they retain. We first describe two negative results. We show that existing proposals to achieve k-*anonymity [23]* in search logs [1], [21], [12], [13] are insufficient in the light of attackers who can actively influence the search log. We then turn to *differential privacy [9]*, a much stronger privacy guarantee; however, we show that it is impossible to achieve good utility with differential privacy.

Algorithm ZEALOUS2, developed independently by Korolova et al. [17] and us [10] with the goal to achieve relaxations of differential privacy. This paper concludes with an extensive experimental evaluation, where we compare the

utility of various algorithms that guarantee anonymity or privacy in search log publishing. Our evaluation includes applications that use search logs for improving both search experience and search performance, and our results show that ZEALOUS' output is sufficient for these applications while achieving strong formal privacy guarantees.

We believe that the results of this research enable search engine companies to make their search log available to researchers without disclosing their users' sensitive information: Search engine companies can apply our algorithm to generate statistics that are (€, δ) probabilistic differentially private while retaining good utility for the applications we have tested. Beyond publishing search logs we believe that our findings are of interest when publishing frequent item sets, as ZEALOUS protects privacy against much stronger attackers than those considered in existing work on privacy preserving publishing of frequent items/item sets [19].

II PRELIMINARIES

The problem of publishing frequent keywords, queries, clicks and other items of a search log are discussed.

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 single user. Such a history is usually partitioned into *sessions* containing similar queries; a *query pair* consists of two subsequent queries from the same user that are contained in 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* analogously. We classify a keyword, query, consecutive query, click in a histogram to be *frequent* if its count exceeds some predefined threshold τ; when we do not want to specify whether we count keywords, queries, etc., we also refer to these objects as *items*.

We define our goal as publishing frequent items (utility) without disclosing sensitive information about the users (privacy).

III.DISCLOSURE LIMITATIONS FOR PUBLISHING SEARCH LOGS

A simple type of disclosure is the identification of a particular user's search history (or parts of the history) in the published search log. The concept of k-anonymity has been introduced to avoid such identifications.

*Definition (*k*-anonymity[23]): A search log is* k *anonymous if the search history of every individual is indistinguishable from the history of at least* k – 1 *other individuals in the published search log.*

UTILITY MEASURE:

We want to publish all items (such as keywords, queries, etc.) with frequency at least τ in a search log; we call such items *frequent items*; we call all other items *infrequent items*. Consider a discrete domain of items D. Each user contributes a set of these items to a search log S. We denote by $f_d(S)$ the frequency of item d Є D in search log S. We drop the dependency from S when it is clear from the context. We define the inaccuracy of a (randomized) algorithm as the expected number of items it gets wrong, i.e., the number of frequent items that are not included in the output, plus the number of infrequent items that are included in the output. We do not expect an algorithm to be perfect. It may make mistakes for items with frequency very close to τ, and thus we do not take these items in our notion of accuracy into account. We formalize this "slack" by a parameter ζ, and given ζ, we introduce the following new notions. We call an item d with frequency $f_d \geq \tau + \zeta$ a *very-frequent item* and an item d with frequency $f_d \leq \tau - \zeta$ a *very-infrequent item*. We will measure the inaccuracy of an algorithm then only using its inability to retain the very-frequent items and its inability to filter out the very infrequent items.

NEGATIVE RESULTS

The deficiency of two existing privacy models for search log publication is discussed. First section focuses on k-anonymity, and second investigates differential privacy.

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. Nevertheless, even without unique identification of a user, an attacker can infer the keywords or queries used by the user.

k- anonymity does not protect against this severe information disclosure.

There is another issue largely overlooked with the current implementations of anonymity. That is instead of guaranteeing that the keywords/queries/sessions of k *individuals* are indistinguishable in a search log they only assure that the keywords/queries/sessions associated with k different *user-IDs* are indistinguishable. These two guarantees are not the same

since individuals can have multiple accounts or share accounts. An attacker can exploit this by creating multiple accounts and submitting the same fake queries from these accounts. It can happen that in a k- keyword/query/session anonymous search log the keywords/queries/sessions of a user are only indistinguishable from k – 1 fake keywords/queries/sessions submitted by an attacker. It is doubtful that this type of indistinguishability at the level of user-IDs is satisfactory.

Privacy Calculation

We introduce a search log publishing algorithm called ZEALOUS that has been independently developed by Korolova et al.and us. ZEALOUS ensures probabilistic differential privacy, and it follows a simple two-phase framework. 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 (referred to as the *sanitized* histogram) is then returned as the output.

IV .PRIVACY–PRESERVING ALGORITHM:

1. For each user u select a set su of up to m distinct items from u's search history.
2. Based on the selected items, create a histogram consisting of pairs (k, ck), where k denotes an item and ck denotes the number of user's u that have k in their search history su. We call this histogram the *original* histogram.
3. Delete from the histogram the pairs (k, ck) with count ck smaller than τ.
4. For each pair (k, ck) in the histogram, sample a random number ηk from the Laplace distribution Lap (λ) 4, and add ηk to the count ck, resulting in a noisy count:

ck ← ck + ηk.

5. Delete from the histogram the pairs (k, ck).
6. Publish the remaining items.

V.CONCLUSIONS

This paper contains a comparative study about publishing frequent keywords, queries, and clicks in search logs. We compare the disclosure limitation guarantees and the

theoretical and practical utility of various approaches. Our comparison includes earlier work on anonymity and (Є', δ')-indistinguishability and our proposed solution to achieve (Є, δ)- probabilistic differential privacy in search logs. In our comparison, we revealed interesting relationships between indistinguishability and probabilistic differential privacy which might be of independent interest. Our results (positive as well as negative) can be applied

more generally to the problem of publishing frequent items or item sets. A topic of future work is the development of algorithms that allow to publish useful information about *infrequent* keywords, queries, and clicks in a search log.

REFERENCES

[1] Eytan Adar. User 4xxxxx9: Anonymizing query logs. In *WWW Workshop on Query Log Analysis*, 2007.
[2] Roberto Baeza-Yates. Web usage mining in search engines. *Web Mining: Applications and Techniques*, 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. In *PODS*, 2007.
[4] Michael Barbaro and Tom Zeller. A face is exposed for aol searcher no. 4417749. New York Times http://www.nytimes.com/2006/08/09/technology/09aol.html?ex=1312776000en=f6f61949c6da4d38ei=5090, 2006.
[5] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *STOC*, pages 609–618, 2008.
[6] Justin Brickell and Vitaly Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *KDD*, 2008.
[7] Soumen Chakrabarti, Rajiv Khanna, Uma Sawant, and Chiru Bhattacharyya. Structured learning for non-smooth ranking losses. In *KDD*, pages 88–96, 2008.
[8] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006.
[9] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and AdamSmith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.
[10] Michaela G¨otz, Ashwin Machanavajjhala, Guozhang Wang, Xiaokui, Xiao, and Johannes Gehrke. Privacy in search logs. *CoRR*, abs/0904.0682v2, 2009.

