ISSN:0975-9646

Publishing Search Logs by Guarantee Users Privacy

Sridevi Sunkavalli[1], A.M.J.Muthukumaran[2]

[1] Student, [2]Asst.prof(Sr.G)
Dept of CSE, SRM University,Chennai

Abstract **- Search engine companies collect the “database of intents”, the stories of their exploiters search queries. To publish search logs with seclusion. These search logs supplies commodious cognition to analyzers and researchers. Search engine fellowships, however, are suspicious of bringing out search logs in order not to divulge sensible information. In this paper we examine algorithms for releasing frequent keywords, queries and clicks of a search log. We foremost show how methods that accomplish variants of k-anonymity are insecure to active attacks. We then show that the harder guarantee ensured by differential secrecy regrettably does not provide any usefulness for this problem. Our paper resolves with a large observational field of study using real diligences where we equivalence ZEALOUS and previous work that attains k-anonymity in search log publishing. Our results show that ZEALOUS generates corresponding utility to k−anonymity while at the same time attaining practically stronger seclusion guarantees.**

Index Terms— Security, integrity, database management, information technology and systems, web search, general.

1 Introduction

When people mention the term "search engine", it is often used generically to describe both crawler-based search engines and human-powered references. In fact, these two types of search engines gather their listings in radically different ways and therefore are inherently different.

A. Crawler-based search engines

Search Engines such as Google, AllTheWeb and AltaVista, create their listings automatically by using a piece of software to “crawl” or “spider” the web and then index what it finds to build the search base. Web page changes can be dynamically caught by crawler-based search engines and will affect how these web pages get listed in the search results.

Crawler-based search engines are good when you have a specific search topic in mind and can be very efficient in finding relevant information in this situation. However, when the search topic is general, crawler-base search engines may return hundreds of thousands of irrelevant responses to simple search requests, including lengthy documents in which your keyword appears only once.

B. **Human-powered directories**

Directories such as the Yahoo directory, Open Directory and LookSmart, depend on human editors to create their listings. Typically, webmasters submit a short description to the directory for their websites, or editors write one for the sites they review, and these manually edited descriptions will form the search base. Therefore, changes made to individual web pages will have no effect on how these pages get listed in the search results.

Human-powered directories are good when you are interested in a general topic of search. In this situation, a directory can guide and help you narrow your search and get refined results. Therefore, search results found in a human-powered directory are usually more relevant to the search topic and more accurate. However, this is not an efficient way to find information when a specific search topic is in mind.

C. Meta-search engines

such as Dogpile, Mamma, and Metacrawler, transmit user-supplied keywords simultaneously to several individual search engines to actually carry out the search.

Search engines play a important role in the sailing through the immensity of the web. Today’s search engines do not just pick up and index webpages, they also pick up and exploit entropy about their exploiters. They depot the queries, clicks, IP-addresses, and other entropy about the interactions with exploiters in what is called a search log. Search logs incorporate valuable entropy that search engines use to tailor their services better to their exploiters needs. They alter the breakthrough of trends, patterns, and anomalies in the search conduct of exploiters, and they can be used in the devastation and testing of new algorithms to meliorate search performance and quality. Scientists all around the world would like to tap this gold mine for their own explore; search engine companies, however, do not exhaust them because they incorporate sensible entropy about their exploiters, for example searches for diseases, lifestyle choices, personal tastes, and political affiliations.

The only exhaust of a search log occurred in 2006 by AOL, and it blended into the annals of tech account as one of the majuscule debacles in the search industry. AOL promulgated three months of search logs of 650,000 exploiters. The only evaluate to assist user privacy was the surrogate of user-ids with ergodic numbers—utterly insufficient security as the New York Times showed by identifying a user from Lilburn, Georgia [4], whose search queries not only incorporated identifying entropy but also sensible entropy about her friends’ ailments. The AOL search log exhaust shows that simply replacing user-ids with ergodic 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 ergodic numbers [18].

In this paper, we equivalence formal methods of limiting disclosure when publishing frequent keywords, queries, and clicks of a search log. The methods 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 seclusion. We then describe Algorithm ZEALOUS,2 evolved independently by Korolova et al. [17] and us [10] with the goal to achieve relaxations of differential seclusion.

2 Preliminaries

In this section, we introduce the problem of publishing frequent keywords, queries, clicks, and other items of a search log.

A. *Search logs*

Search engines such as Bing, Google, or Yahoo log interactions with their exploiters. 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 incorporating similar queries; how this partitioning is done is orthogonal to the techniques in this paper. A query pair consists of two subsequent queries from the same user within the same session.

B. *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 1 (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.

There are several proposals in the literature to achieve different variants of k-anonymity for search logs. Adar proposes to partition the search log into 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 ergodic number [1]. We call the output of Adar's Algorithm a k-query anonymous search log. Motwani and Nabar add or delete keywords from sessions until each session incorporates the same keywords as at least k -1 other sessions in the search log [21], following by a replacement of the user-id by a ergodic number. We call the output of this algorithm a k-session anonymous search log. He and 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 [12]. We call the output a k-keyword anonymous search log. Efficient ways to anonymize a search log are also discussed by Yuan et al. [13].

Definition 2 (e - differential privacy [9]). An algorithm A is e- differentially private if for all search logs S and S^1 differing in the search history of a single user and for all output search logs O:

$$\Pr [A(S)]= O] <= e^{e} \Pr [A(S^1)= O].$$

This definition ensures that the output of the algorithm is insensitive to changing/omitting the complete search history of a single user. We will refer to search logs that only differ in the search history of a single user as neighboring search logs. Similar to the variants of k-anonymity, we could also define variants of differential seclusion by looking at neighboring search logs that differ only in the content of one session, one query or one keyword. However, we chose to focus on the strongest definition in which an attacker learns roughly the same about a user even if that user's whole search history was omitted.

Existing work on publishing frequent itemsets often only tries to achieve anonymity or makes strong assumptions about the background knowledge of an attacker. The main focus of this paper is search logs, our results apply to other scenarios as well. For example, consider a retailer who collects customer transactions. Each transaction consists of a basket of products together with their prices, and a time-stamp. In this case ZEALOUS can be applied to publish frequently purchased products or sets of products. This information can also be used in a recommender system or in a market basket analysis to decide on the goods and promotions in a store. Our results show that ZEALOUS yields comparable utility to k–anonymity while at the same time achieving much stronger seclusion guarantees.

C. *Utility Measures*

i) Theoretical Utility Measures

For simplicity, suppose we want to publish all items (such as keywords, queries, etc.) with frequency at least *T* 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 logS. We denote by $f_d(S)$ the frequency of item d € D in search log S. We drop the dependency from S when it is clear from the context.

ii) Experimental Utility Measures

Traditionally, the utility of a privacy-preserving algorithm has been evaluated by comparing some statistics of the input with the output to see "how much information is lost." The choice of suitable statistics is a difficult problem as these statistics need to mirror the sufficient statistics of applications that will use the sanitized search log, and for some applications the sufficient statistics are hard to characterize. To avoid this drawback, Brickell and Shmatikov [6] measure the utility with respect to data mining tasks and they take the actual classification error of an induced classifier as their utility metric.

In this paper, we take a similar approach. We use two real applications from the information retrieval community: Index caching, as a representative application for search performance, and query substitution, as a representative application for search quality. For both application, the sufficient statistics are histograms of keywords, queries, or query pairs.

3 Negative Results

A. *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.

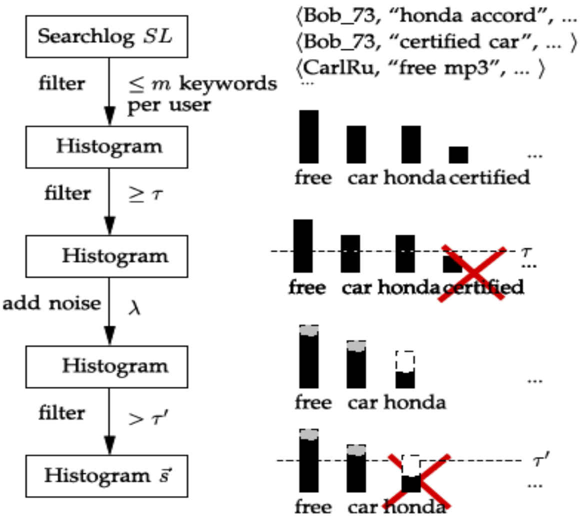
B. *Impossibility of Differential Privacy*

In the following, we illustrate the infeasibility of differential seclusion in search log publication. In particular, we show that, under realistic settings, no differentially private algorithm can produce a sanitized search log with reasonable utility. Our analysis is based on the following lemma.

Lemma: For a set of U users, let S and S0 be two search logs each containing at most m items from some domain D per user. Let A be an e-differentially private algorithm that, given S, retains a very frequent item d in S with probability p. Then, given S0, A retains d with probability at least $p/(e^{L1(S,S0)})^{-e/m})$, where $L1(S, S0) = \sum_{d€D}|f(s) = f(s0)|$ denotes the L1 distance between S and S0.

4 Achieving Privacy

Privacy preserving algorithm



Algorithm ZEALOUS for Publishing Frequent Items of a Search Log

5 Beyond Search Logs

While the main focus of this paper are search logs, our results apply to other scenarios as well. For example, consider a retailer who collects customer transactions. Each transaction consists of a basket of products together with their prices, and a time stamp. In this case, ZEALOUS can be applied to publish frequently purchased products or sets of products. This information can also be used in a recommender system or in a market basket analysis to decide on the goods and promotions in a store [11]. Another example concerns monitoring the health of patients. Each time a patient sees a doctor, the doctor records the diseases of the patient and the suggested treatment. It would be interesting to publish frequent combinations of diseases.

All of our results apply to the more general problem of publishing frequent items/item sets/consecutive itemsets. Existing work on publishing frequent item sets often only tries to achieve anonymity or makes strong assumptions about the background knowledge of an attacker, see, for example, some of the references in the survey by Luo et al. [19].

6 Conclusions

This paper incorporates a comparative study about issuing frequent keywords, queries, and clicks in search logs. We equivalence the disclosure limitation ensures and the theoretical and practical utility of various approaches. Our comparison includes earlier work on anonymity and (e^1, ð^1)-indistinguishability and our proposed solution to achieve (e,ð)- probabilistic differential seclusion in search logs. In our comparison, we revealed interesting relationships between indistinguishability and probabilistic differential seclusion 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 to exhaust useful information about infrequent keywords, queries, and clicks in a search log while preserving user privateness.

References

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

[2] R. Baeza-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 SIGMODSIGACT- 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?ex=1312776000en= f6f61949c6da4d38ei=5090, 2006.

[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.

[7] S. Chakrabarti, R. Khanna, U. Sawant, and C. Bhattacharyya, "Structured Learning for Non-Smooth Ranking Losses," Proc. ACM

SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 88-96, 2008.

[8] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our Data Ourselves: Privacy via Distributed Noise Generation," Proc. Ann. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), 2006.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," Proc. Theory of Cryptography Conf. (TCC), 2006.

[10] M. Go¨ tz, A. Machanavajjhala, G. Wang, X. Xiao, and J. Gehrke, "Privacy in Search Logs," CoRR, abs/0904.0682v2, 2009.

[11] J. Han and M. Kamber, Data Mining: Concepts and Techniques, first ed. Morgan Kaufmann, Sept. 2000.

[12] Y. He and J.F. Naughton, "Anonymization of Set-Valued Data via Top-Down, Local Generalization," Proc. VLDB Endowment, vol. 2, no. 1, pp. 934-945, 2009.

[13] Y. Hong, X. He, J. Vaidya, N. Adam, and V. Atluri, "Effective Anonymization of Query Logs," Proc. ACM Conf. Information and Knowledge Management (CIKM), 2009.

[14] R. Jones, R. Kumar, B. Pang, and A. Tomkins, "I Know What You Did Last Summer: Query Logs and User Privacy," Proc. ACM Conf. Information and Knowledge Management (CIKM), 2007.

[15] R. Jones, B. Rey, O. Madani, and W. Greiner, "Generating Query Substitutions," Proc. 15th Int'l Conf. World Wide Web (WWW), 2006.

[16] S. Prasad Kasiviswanathan, H.K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What Can We Learn Privately?" Proc. 49th Ann. IEEE Symp. Foundation of Computer Science (FOCS), pp. 531-540, 2008.

[17] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas, "Releasing Search Queries and Clicks Privately," Proc. 18th Int'l Conf. World Wide Web (WWW), 2009.

[18] R. Kumar, J. Novak, B. Pang, and A. Tomkins, "On Anonymizing Query Logs via Token-Based Hashing," Proc. Int'l Conf. World Wide Web (WWW), 2007.

[19] Y. Luo, Y. Zhao, and J. Le, "A Survey on the Privacy Preserving Algorithm of Association Rule Mining," Proc. Int'l Symp. Electronic Commerce and Security, vol. 1, pp. 241-245, 2009.

[20] A. Machanavajjhala, D. Kifer, J.M. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory Meets Practice on the Map," Proc. Int'l Conf. Data Eng. (ICDE), 2008.

[21] R. Motwani and S. Nabar, "Anonymizing Unstructured Data," Corr, abs/0810.5582, 2008.

[22] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth Sensitivity and Sampling in Private Data Analysis," Proc. Ann. ACM Symp. Theory of Computing (STOC), 2007.

[23] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.