Information Sciences 460–461 (2018) 223–237

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A semantic-preserving differentially private method for releasing query logs

David Sánchez[a,*], Montserrat Batet[b], Alexandre Viejo[a], Mercedes Rodríguez-García[c], Jordi Castellà-Roca[a]

[a] Department of Computer Science and Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia, UNESCO Chair in Data Privacy, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Spain

[b] Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Av. Carl Friedrich Gauss, 5, Parc Mediterrani de la Tecnologia, 08860 Castelldefels (Barcelona), Spain

[c] School of Engineering, University of Cadiz, Av. de la Universidad 10, 11519 Puerto Real, Cadiz, Spain

ARTICLE INFO

Article history:
Received 17 March 2017
Revised 17 May 2018
Accepted 21 May 2018
Available online 28 May 2018

Keywords:
Differential privacy
Query logs
Data utility
User profiling

ABSTRACT

Query logs are of great interest for data analysis. They allow characterizing user profiles, user behaviors and search habits. However, since query logs usually contain personal information, data controllers should implement appropriate data protection mechanisms before releasing them for secondary use. In the past, the anonymization of query logs was tackled from the perspective of statistical disclosure control and by relying on privacy models such as k-anonymity, which do not scale well with the high dimensionality and dynamicity of query logs. To offer better privacy protection, some authors have recently embraced the robust privacy guarantees of ε-differential privacy. However, this comes at the cost of limiting the number and types of analyses that can be made on the protected queries. To tackle this issue, in this paper we propose a privacy protection method for query logs that joins the flexibility and convenience of privacy-preserving data releases with the strong privacy guarantees of ε-differential privacy. Moreover, to retain the analytical utility of the protected query, we have put special care in capturing, managing and preserving the semantics of the queries during the protection process. The empirical experiments we report show that our method produces differentially private query logs that are more useful for analysis than related works.

1. Introduction

Among the activities performed by Internet users, the most common action is probably the use of web search engines (WSEs) as the entry point for finding information and surfing around the Web. When a WSE receives a search query, it looks for the required information among billions of indexed web pages, and returns the corresponding search results in the form of ranked documents. During this process, the WSE automatically stores the submitted query (i.e., the keywords) and some related metadata (e.g., date of the query, an identifier of the sender or the search result selected by the sender). The recorded search queries together with their metadata are stored in files named *query logs* [9].

The analysis of these query logs allows the characterization of user profiles, user behaviors and search habits. This information represents the cornerstone of marketing techniques such as *Behavioral Targeting*, which are applied by website

[*] Corresponding author.
E-mail address: david.sanchez@urv.cat (D. Sánchez).

https://doi.org/10.1016/j.ins.2018.05.046

publishers and advertisers to increase the effectiveness of advertisements [10]; or the so-called *Search Engine Marketing*, which employs user profiles and related search data to improve keyword advertising campaigns and extract market tendencies, among others [21]. Marketing companies and other third parties are well aware of the usefulness of query logs and they buy these data from WSE operators in order to exploit them for their own economic purposes [35]; specifically, constructing user profiles by means of the data stored in query logs has been acknowledged as a relevant topic by the scientific community and has received significant attention [20,46].

Despite its benefits, the exploitation of query logs does not come without cost. In particular, query logs may contain:

(i) Pieces of data that may allow re-identifying the individuals who have generated them; for instance, queries related to the user's address, sex, occupation or age may be aggregated to enable univocal re-identifications [43].
(ii) Information related to very sensitive topics, such as religion, sexual preferences or medical conditions, among others [3]; these may be linked to individuals in case of re-identification and may be used for discriminatory purposes (e.g., in health insurance or credit applications) [35].

The uncontrolled disclosure of query logs, thus, poses a serious privacy threat to the users of WSEs. In order to mitigate this threat, query logs should be anonymized before releasing them to untrusted parties for secondary use. The anonymization method should minimize the probability of re-identifying individuals, while ensuring that the protected data are still useful for analysis; specifically, this would allow data controllers to generate anonymous but accurate enough user profiles that may still be exploited for a variety of purposes.

1.1. Related work

The anonymization of query logs has already been tackled in the literature from different perspectives.

The most straightforward schemes use two kind of approaches:

(i) Removal of queries that are considered to be specially privacy-threatening [8], such as *infrequent queries* [1] or queries that contain *direct identifiers* of the individuals (e.g., social security numbers, names or addresses) [3].
(ii) Mix of queries among senders that share the same interests in a way that the users never appear linked to their own original queries in the anonymized query logs [29].

These techniques do not rely on privacy models, which constitute the means to provide formal and beforehand privacy guarantees on the protected query logs. As a result, they fail to balance the trade-off between privacy protection and data utility preservation, which is the main goal behind privacy-preserving data releases.

Among the systems enforcing privacy models, those based on *k-anonymity* [36] are the most common ones [4,7,28]. They build groups of k users and aggregate or generalize their query logs in order to make them indistinguishable and, thus, lower the probability of re-identifying the individuals to, at most, $1/k$. Nevertheless, k-anonymity was designed for structured databases with a limited number of attributes and, therefore, its performance is severely hampered with high dimensional data such as query logs [2]. Moreover, because k-anonymity relies on making data sets more uniform, it is designed to work with static data. As a result, it does not easily support privacy-preserving updates, either adding more queries to an existing user or adding new users' logs.

ε-Differential privacy, on the other hand, is a privacy model that was originally proposed in [14] to be used in *interactive settings*. In these settings, an anonymization mechanism sits between an analyst submitting (statistical) queries on the sensitive data set and a trusted database holder that stores the clear data and provides sanitized (differentially private) answers. *ε-Differential privacy* guarantees that the protected response to a certain statistical query is very similar to the output obtained for the same query when the data of an individual are removed (or modified) from the data set. In this way, the presence or absence of the data of a certain individual in the data set does not have a significant influence on the protected results, thus, preventing attackers from performing re-identification inferences. This privacy guarantee is stronger than the one provided by k-anonymity; however, it comes at the cost of severely restricting the type, number and/or accuracy of the statistical queries submitted to the database.

Some researchers have used ε-differential privacy to protect query logs. In [22] the authors define an interactive scenario in which the entities that send requests to a remote database holding the query logs are classified according to a certain "requester profile". Once a requester is assigned to a certain profile, the proposed scheme applies a privacy policy to the requester, which is based on adding a certain level of noise to the answers in order to fulfill differential privacy. The goal of this proposal is to enable stakeholders to perform general privacy-preserving data analyses; that is, it is capable of answering statistical queries such as "number of people searching for Starbucks over a period of one month", but it cannot answer to individual-level queries.

Other approaches are based on releasing a differentially private data structure (e.g., a query click graph), in which the queries performed by the users and their corresponding clicks are aggregated without their individual attribution. Proposals that follow this approach generate a protected output that keeps intact the statistics needed to implement the typical uses of query logs, such as query suggestions, inferring spelling corrections or classifying queries according to a set of topics [23,27]; or to compute common information retrieval operations such as "ordering of top results" or "re-ranking of query results" [49,50]. However, by discarding the users that have performed the queries, these methods cannot be used to build user profiles.