Knowledge-Based Systems 101 (2016) 71–89

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Hierarchical anonymization algorithms against background knowledge attack in data releasing

Fatemeh Amiri[a], Nasser Yazdani[a], Azadeh Shakery[a,b,*], Amir H. Chinaei[c]

[a] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
[b] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), P.O. Box 19395-5746 Tehran, Iran
[c] Department of Electrical and Computer Engineering, University of Puerto Rico at Mayaguez, Mayaguez, PR, USA

ARTICLE INFO

Article history:
Received 15 August 2015
Revised 5 March 2016
Accepted 7 March 2016
Available online 12 March 2016

Keywords:
Privacy preservation
Tabular data
Hierarchical anonymization algorithm
Background knowledge
Information loss metric

ABSTRACT

Preserving privacy in the presence of adversary's background knowledge is very important in data publishing. The k-anonymity model, while protecting identity, does not protect against attribute disclosure. One of strong refinements of k-anonymity, β-likeness, does not protect against identity disclosure. Neither model protects against attacks featured by background knowledge. This research proposes two approaches for generating k-anonymous β-likeness datasets that protect against identity and attribute disclosures and prevent attacks featured by any data correlations between QIs and sensitive attribute values as the adversary's background knowledge. In particular, two hierarchical anonymization algorithms are proposed. Both algorithms apply agglomerative clustering techniques in their first stage in order to generate clusters of records whose probability distributions extracted by background knowledge are similar. In the next phase, k-anonymity and β-likeness are enforced in order to prevent identity and attribute disclosures. Our extensive experiments demonstrate that the proposed algorithms outperform other state-of-the-art anonymization algorithms in terms of privacy and data utility where the number of unpublished records in our algorithms is less than that of the others. As well-known information loss metrics fail to measure precisely the imposed data inaccuracies stemmed from the removal of records that cannot be published in any equivalence class. This research also introduces an extension into the Global Certainty Penalty metric that considers unpublished records.

1. Introduction

Advances in the Internet and data processing technologies have accelerated data collection and dissemination. As collected data may contain private information, a breach of privacy is possible if it is disclosed-together with identifiers-to unauthorized parties. Removal of attributes that are identifiers, such as name and social security number, is not sufficient to protect privacy, when quasi identifiers[1] (QI) exist. Hence, proposing promising approaches for privacy preservation has gained significant attention in the context of data collection and dissemination.

Anonymization is an approach to preserve individuals' privacy by removing their identifiers from the data that is going to be published, while maintaining as much of original information as possible. Each anonymization framework includes a privacy model and an anonymization algorithm. Privacy models can be divided into *syntactic* and *semantic* models. Syntactic privacy models partition data into a set of groups (called equivalence classes) such that all records within each equivalence class are indistinguishable from one another from QI point of view. In the k-anonymity model, as the first syntactic privacy model, each equivalence class contains at least k records [1,2]. This model prevents identity disclosure[2], but it does not preserve the privacy against attribute disclosure[3]. To address this issue, other variants of k-anonymity have been proposed [3–5]. Semantic privacy models add some noise to data in order to preserve privacy. The *differential privacy model* is a semantic privacy model in which it is guaranteed that deletion and addition of any individual's record does not significantly affect the result of data analysis [6].

* Corresponding author at: School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran. Tel.: +982182089722.

E-mail addresses: f.amiri@ece.ut.ac.ir (F. Amiri), yazdani@ut.ac.ir (N. Yazdani), shakery@ut.ac.ir (A. Shakery), Amirhossein.Chinaei@ece.uprm.edu (A.H. Chinaei).

[1] Quasi identifiers are those attributes that individually are not an identifier, but when combined together, they become identifiers.

[2] The individual, to whom a record is associated, cannot successfully be re-identified with probability more than $\frac{1}{k}$.

[3] Attribute disclosure occurs when sensitive information about an individual is revealed.

http://dx.doi.org/10.1016/j.knosys.2016.03.004
0950-7051/© 2016 Elsevier B.V. All rights reserved.

Each privacy model provides a defense against a particular adversary model. A common assumption is that the adversary has two pieces of information: (I) whether or not his/her targets exist in the microdata table and (II) the QI values of his/her targets [1–4]. None of the models mentioned above- including the differential privacy model- can preserve privacy if the adversary has additional information (called background knowledge) [7]. Hence, researchers have proposed enhanced models that assume the adversary has some background knowledge [8–13]. A background knowledge is any known fact that by itself is not a privacy disclosure, but the adversary combines it with other information to make more precise inference on target's sensitive information. This is called a background knowledge attack. Examples of background knowledge in a particular medical dataset context are "a male breast cancer is rare", "the prevalence of chronic bronchitis is higher among the 65+ age group compared to other groups; and, across all age groups, females have higher rates than males for both black and white races", etc. [14].

In this work, we develop a syntactic-based anonymization framework in which we assume the adversary has background knowledge about the correlations among dataset attributes. In a syntactic privacy model, when an equivalence class is published, the adversary can estimate the probabilities of possible associations of sensitive values to his/her target (i.e. record respondent) without exploiting any background knowledge[4]. When different sensitive attribute values exist in an equivalence class, the probability of associating a record respondent to sensitive values in the equivalence class is the same. By exploiting the background knowledge, the adversary may be able to discriminate one association from the others, resulting in privacy breaches. Modeling the background knowledge is an open problem in data anonymization [8]. We model the adversary's background knowledge as a probability distribution associating the sensitive values to a record respondent based on QI values, called *background knowledge distribution*. The goal in our privacy model is to maximize uncertainties in identifying record respondents and their respective values for sensitive attributes in a given equivalence class. In the presence of adversary's background knowledge, we attempt to create equivalence classes such that record respondents have similar *background knowledge distributions* in each class in order to achieve our goal. Therefore, when adversaries examine different associations of sensitive values (within each equivalence class) to their targets, they will not be able to discriminate any association with a high degree of certainty. The constraint of similar *background knowledge distributions* cannot prevent identity disclosure or attribute disclosure. Therefore, we also apply k-anonymity [1] and β-likeness [5]. To remain the anonymized data useful, a high similarity among QI values in each equivalence class is also required.

Hence, we propose to create equivalence classes with the following privacy requirements: (1) The *background knowledge distributions* within any equivalence class should be similar in order to prevent the so-called background knowledge attack, (2) k-anonymity: the size of each class is at least k, (3) β-likeness: the maximum relative difference in the frequency of sensitive values within any equivalence class and that of the overall microdata table does not exceed a given threshold β.

We present two syntactic anonymization algorithms based on the value generalization approach. We suggest a hierarchical procedure to satisfy our privacy requirements. First, we apply agglomerative clustering to prevent the background knowledge attack. We apply the clustering algorithm to generate clusters in which the difference of *background knowledge distributions* between each pair of records in the cluster is below a certain threshold. Then, each cluster is partitioned into a number of equivalence classes. We propose two algorithms to produce the equivalence classes: k-anonymity-primacy and β-likeness-primacy. The former prioritizes the QI attributes and generates equivalence classes in a β-likeness aware manner. For this purpose, we propose a clustering-based algorithm to select homogeneous records in terms of QI values, and then check whether β-likeness is satisfied. The latter focuses on the sensitive attribute values. It generates large equivalence classes in which β-likeness is satisfied. Then, large equivalence classes are split in order to satisfy k-anonymity.

A work close to ours is found in Riboni et al. [8]. They have proposed a privacy model based on adversary's background knowledge and t-closeness [4]. Their anonymization algorithm applies Hilbert index transformation to create an ordered list of record respondents based on similarity of QI-values. We enforce a stronger model than t-closeness and propose two new anonymization algorithms to satisfy our privacy requirements. Sori-Comas et al. [15] also proposed two clustering-based anonymization algorithms attaining k-anonymity and t-closeness which is suitable to anonymize numerical values. They do not consider the adversary's background knowledge.

We verify the effectiveness of our anonymization algorithms by running extensive experiments on two datasets: Adult dataset [16] and BKseq dataset [8]. We study the performance of our anonymization algorithms based on different parameters of our privacy model. The experimental results show that k-anonymity-primacy generates the anonymized microdata with low information loss while β-likeness-primacy incurs low privacy loss. The k-anonymity-primacy algorithm also generates more balanced equivalence classes compared to β-likeness-primacy. We further compare the performance of our proposed algorithms with the state of the art anonymization algorithms like Hilbert index-based algorithm [8]. The performance of our algorithms are better than Hilbert index-based algorithm in terms of both data utility and privacy.

Furthermore, we extend an information loss measure to capture data inaccuracies caused by generalization. In any anonymization algorithm, it is possible not to fit some records in any equivalence class. To protect the privacy of other record respondents, a simple solution is to remove them from the published data. We introduce an extension to the Global Certainty Penalty (GCP) metric [17] to consider this kind of information loss too, and we name it *Removed Global Certainty Penalty* (RGCP). The RGCP metric charges a penalty for each not-fit record. The penalty of each record is proportional to the range of QI values in the nearest equivalence class. We evaluate our algorithms using both GCP and RGCP. When a large number of records are removed by the algorithm, the advantage of RGCP over GCP is better seen.

Contributions. In summary, our contributions are as follows:

- We propose two syntactic anonymization algorithms which simultaneously satisfy two privacy models (k-anonymity and β-likeness) against adversaries who have background knowledge on correlations of attributes. We conduct extensive experiments on different aspects of the algorithms, namely data utility, privacy, size of equivalence classes, and the run time. Then we compare our algorithms with microaggregation approaches. We also perform an experimental comparison between the closest work, Hilbert index-based algorithm [8], and the proposed algorithms. We demonstrate that the proposed algorithms outperform Hilbert index-based algorithm in terms of data utility and privacy.
- We extend GCP to measure information loss of equivalence classes when the generalization operation is performed. Our metric, called RGCP, considers unpublished records too.

[4] Recall that adversaries know the QI values of their targets.