Privacy Preserving Data Publishing - Survey

Shumin Guo

Summary

This paper has made a comprehensive survey of privacy preserving data publishing(PPDP) research. It first defines various privacy scenarios for data publishing privacy protection, and discusses current available privacy and utility metrics, analyzes deterministic and randomized sanitization techniques and algorithms. Then this paper identifies ways about how to use sanitized data by legitimate users and misuse by attackers. Finally, it summarizes privacy issues that arises in social networks and pervasive computing environments and concludes this paper by identifying challenges for research on PPDP.

Privacy Definitions

Privacy issue arises when data owners(publishers) want to publish data, and on the other hand want to protect the privacy of individual users. PPDP is trying to tackle this problem.

An intuitive method to protect privacy is to anonymize the real data by removing identifiers to hide the sensitive information. But due to the possibility that adversaries may have background knowledge or other sources of public information, this method can not properly protect the privacy. K-anonymity is proposed to avoid this linking attack on anonymized data. This methods tries to guarantee that quasi-identifiers(QIs) occurs at least k times in a specific data set.

ℓ-diversity privacy is proposed in case that adversaries have more background knowledge and can infer sensitive information about individuals even without re-identifying them. Examples are *homogeneity attack* and *background knowledge attack*. This privacy method is to make sure that sensitive values/attributes have enough diversity and adversaries can not uniquely identify sensitive values associated with a group. *Recursive (c, ℓ)-diversity* is an instantiation of the ℓ-diversity method.

A more general way to represent background knowledge attack is to describe background knowledge as boolean logic sentences and try to protect adversaries who might know a certain number of such sentences. With this framework, ℓ-diversity and k-anonymity are just special cases.

Other than precise background knowledge attacks, adversaries may obtain knowledge with probability. (α-β)-privacy is proposed for probabilistic background knowledge, this privacy considers all possible inputs and outputs of an anonymization algorithm. And it does not limit the information known to adversaries by considering all possible prior knowledge owned by adversaries. So, no deterministic algorithms can satisfy the (α-β)-privacy, and we need to consider random perturbation based methods.

Differential privacy is proposed to make sure that an individual's privacy is guaranteed if given access to the sanitized data set and information about all but on individual in a table, an adversary can not determine the value of this individual's tuple. It considers the safety of an anonymization algorithm over all possible inputs and outputs.

Perfect privacy is used in cases when an individual doesn't want to disclosure any personal information.

Except for the privacy definitions above, there are various extensions and relaxations for these privacy. In summary, all the privacy definitions are based on the assumption that adversaries have different background knowledge. But how to select a privacy method in a specific situation is yet to be answered. In order to answer this question, a generalized privacy framework is proposed to consider three important questions: what private information to protect, how is private information leaked and when disclosed private information lead to privacy breach and how to measure this disclosure. Based on this, we can put all the above privacy definitions into a universal framework.

Utility metrics

Various methods have been proposed to measure the utility of sanitized data. These methods include generalization/suppression counting which measures utility by counting the number of anonymization operations over a data set, Loss metric which is defined in terms of a normalized loss for each attribute of every tuple, classification metric which is defined by measuring the effect of the sanitization on a hypothetical classifier, discernibility metric which is a panelty based metric, ambiguity based metric which is based on the ambiguity of tuples in sanitized data, KL-divergence which is based on a statistical method, L_p norm which is used to measure the distance between the original and reconstructed probability distributions, hellinger distance, bivariate measures, work-load aware metric, first and second order statistics, analytical, invariance and reconstructibility.

Sanitization Algorithms

Data sanitization algorithm includes two big categories, one is deterministic which doesn't introduce randomness into the process and another is random sanitization algorithms which use random perturbation methods.

Examples of deterministic sanitization algorithms are suppression based algorithms which hides microdata by suppression, generalization algorithms which generalizes values into a number of intermediate states by recoding and structured aggregation, microaggregation which involves data partitioning and partition aggregation steps, bucketization which is used to achieve ℓ-diversity by partitioning and breaking the connections between QIs and sensitive attributes, decomposition and marginals which selects subsets of attributes to produce low-dimensional data table or project a single table of microarray data on selected attribute subsets.

Randomized data sanitization methods include local randomization which is used to randomly perturb one or a group of sensitive attributes so that the privacy of individuals can be protected at the data collection stage, input randomization which perturbs data before the input stage and additive perturbation and post randomization are such techniques, statistics perturbation which perturbs the statistical distribution of the sensitive data set and can guarantee differential privacy, statistics preserving input randomization, model based synthetic data generalization in which a statistical model is generated from a noise infused version of the existing data and sample data points from this model to create synthetic data set.

Using Sanitized Data

Sanitized data can be used in various ways. Examples are data querying such as probabilistic querying which is about querying data over a probablistic database, OLAP over uncertain and imprecise data, machine learning and data mining algorithms over the sanitized data sets and statistical analysis of the sanitized data set. Also if the data are used by a malicious user, various attacks maybe possible. Examples of attacks are combinational attack which is about attack based on certain specific background knowledge of the victim(s) over the database, optimality attack in which the attacker only knows the non-sensitive information along with the privacy policy and algorithms used for anonymization, alternative reasoning attack which reasons about the sanitized data in a way that is different from the way used by data publishers and thus can poke information from data sets, denoising attack which tries to remove noise from sanitized data, undesired use of data in which data is used in and undesired way from the data publishers' point of view and thus have the possiblity of cause privacy breach. Various attacks also consider using external information such as linking attacks, composition attack, similar data attack, instance level background knowledge attacks.

Challenges and New Applications of PPDP

Various new applications, decentralized data collection and pervasive computing bring new challenges to the area of PPDP. Social networking privacy is one such example. Location privacy related to mobile computing is another exmaple.