Slicing a New Method for Data Publishing Preserving the Privacy

R.Kathyayani*1, S.G.Nawaz*2*2

M.Tech, Dept of CSE, SKDEC, Gooty, D.t: Anantapuram, A.P, India

M.Tech, Associate Professor, Dept of CSE, SKDEC, Gooty, D.t: Anantapuram, A.P, India

ABSTRACT

Privacy is an important issue in data publishing. Many organizations distribute non-aggregate personal data for research, and they must take steps to ensure that an adversary cannot predict sensitive information pertaining to individuals with high confidence. This problem is further complicated by the fact that, in addition to the published data, the adversary may also have access to other resources (e.g., public records and social networks relating individuals), which we call *external knowledge*. A robust privacy criterion should take this external knowledge into consideration. In this paper, we ask whether generalization and suppression of quasi-identifiers offer any benefits over trivial sanitization which simply separates quasi-identifiers from sensitive attributes. Previous work showed that k- anonymous databases can be useful for data mining, but k-anonymization does not guarantee any privacy. By contrast, we measure the tradeoff between privacy (how much can the adversary learn from the sanitized records?) and utility, measured as accuracy of data-mining algorithms executed on the same sanitized records.

KEYWORDS: quasi identifiers, data publishing, privacy.

I.INTRODUCTION

A number of recent high-profile attacks have illustrated the importance of protecting individuals' privacy when publishing or distributing sensitive personal data. For example, by combining a public voter registration list and a released database of health insurance information, Sweeney was able to identify the medical record of the governor of Massachusetts. In the context of data publishing, it is intuitive to think of privacy as a game between a data owner, who wants to release data for research, and an adversary, who wants to discover sensitive information about the individuals in the database. Following most of the previous literature, we take a constrained optimization approach. That is, the data owner seeks to find the "snapshot" (*release candidate*) of her original dataset that simultaneously satisfies the given privacy criterion and maximizes some utility measure. Note that the privacy criterion determines the safety of the released data, and the utility measure is an orthogonal issue. The focus of this paper is developing a practical privacy criterion that captures the problem of attribute disclosure in the presence of external knowledge. Specifically, we consider the case where the data owner has a table of data (denoted by **D**), in which each row is a record pertaining to some individual. The attributes of this table consist of (1) a set of *identifier* (ID) attributes which will be removed from the released dataset, (2) a set of *quasi-identifier* (QI) attributes that together can potentially be used to re-identify individuals, and (3) a *sensitive* attribute (denoted by *S*), which is possibly set-valued. For example, consider the original data in Figure 1. In this example, *Name* is the ID attribute. *Age*, *Gender*, *Zipcode* are the QI attributes, and *Disease* is the sensitive attribute. Microdata records contain information about specific individuals. Examples include medical records used in public-health research, individual transactions or preferences re-leased to support the development of new data-mining algorithms, and records published to satisfy legal requirements.

In contrast to statistical databases and randomized response methods, the records in question contain actual, unperturbed data associated with individuals. Some of the attributes may be sensitive, e.g., health-related attributes in medical records. Therefore, identifying attributes such as names and Social

Security numbers are typically removed from microdata records prior to release. The published records may still contain quasi-identifiers," e.g., demo- graphic attributes such as ZIP code, age, or sex. Even though the quasi-identifier attributes do not directly reveal a person's identity, they may appear together with the identity in another public database, or it may be easy to reconstruct their values for any given individual. Microdata records may also contain \neutral" attributes which are neither quasi-identifying, nor sensitive. The association of quasi-identifiers with sensitive attributes in public records has long been recognized as a privacy risk. This type of privacy breach is known as sensitive attribute disclosure, and is different from membership disclosure, i.e., learning whether a certain individual is included in the database. It is very easy to prevent sensitive attribute disclosure by simply not publishing quasi-identifiers and sensitive at- tributes together. Trivial sanitization that removes either all quasi-identifiers or all sensitive attributes in each data release provides the maximum privacy possible against an ad- adversary whose knowledge about specific individuals is limited to their quasi-identifiers (this adversary is very weak, yet standard in the microdata sanitization literature.

There is large body of research on techniques such as k- anonymity and `-diversity that apply domain-specific generalization and suppression to quasi-identifier attributes and then publish them together with unmodified sensitive attributes. In this paper, we ask a basic question: what benefit do these algorithms provide over trivial sanitization? The only reason to publish generalized quasi-identifiers and sensitive attributes together is to support data-mining tasks that consider both types of attributes in the sanitized database. Our goal in this paper is to evaluate the tradeoff between this incremental gain in data-mining utility and the degradation in privacy caused by publishing quasi-identifiers together with sensitive attributes. Our contributions. First, we give a semantic definition of sensitive attribute disclosure. It captures the gain in the adversary's knowledge due to his observations of the sanitized dataset. This definition is somewhat similar to privacy definitions used in random-perturbation databases, but is adapted to the generalization and suppression framework.

Second, we give a methodology for measuring the tradeoff between the loss of privacy and the gain of utility. Privacy loss is the increase in the adversary's ability to learn sensitive attributes corresponding to a given identity. Utility gain is the increase in the accuracy of machine-learning tasks evaluated on the sanitized dataset. The baseline for both is the trivially sanitized dataset, which simply omits either all quasi-identifiers, or all sensitive attributes, thus providing maximum privacy and minimum utility. Third, we evaluate our methodology on the same datasets from the UCI machine learning repository as used in previous research on sanitized microdata utility. We show that non-trivial generalization and suppression either results in large privacy breaches, or provides little incremental utility vs. a trivially sanitized dataset. Therefore, even if the adversary's knowledge is limited to quasi-identifiers, the data-mining utility must be destroyed to achieve only marginal privacy. To protect against an adversary with auxiliary knowledge, the loss of utility must be even greater.

II.RELATED WORK

Privacy in statistical databases has been a topic of much research. Techniques include adding random noise to the data while preserving certain statistical aggregates and interactive output perturbation .

By contrast, microdata publishing involves releasing unperturbed records containing information about individuals. k-anonymity is a popular interpretation of privacy. Many methods have been proposed for achieving it most apply generalization and suppression to quasi-identifiers only. In Section 6, we compare our experimental methodology to previous work. Limitations of k-anonymity are: (1) it does not hide whether a given individual is in the database [26, 30], (2) it reveals individuals' sensitive attributes [21, 22], (3) it does not protect against attacks based on background knowledge [22, 23], (4) mere knowledge of the k-anonymization algorithm can violate privacy [43], (5) it cannot be applied to high-dimensional data without complete loss of utility [3], and (6) special methods are required if a dataset is anonymized and published more than once .

In, fihrn and Ohno-Machado proposed that the sensitive attributes associated with each quasi-identifier be diverse." This is similar to p-sensitivity, `-diversity and others. Diversity of sensitive attributes, however, is neither necessary, nor sufficient to prevent sensitive attribute disclosure. A stronger definition appears in [24], but it is unclear whether it can be achieved in the data access model considered in the generalization and suppression framework.

In the k-anonymity literature, the adversary's knowledge is limited to quasi-identifiers such as age and ZIP code.

Stronger adversaries with background knowledge are considered in [9, 23]. Our results show that generalization and suppression do not protect privacy even against very weak adversaries who only know the quasi-identifiers; privacy obviously fails against stronger adversaries as well.

This paper is about sensitive attribute disclosure. Membership disclosure, i.e., learning whether a given individual is present in the sanitized database, is a different, incomparable privacy property. Methods for preventing membership disclosure such as [12,26,30] are complementary to our work.

III. Database Privacy

Intuitively, to achieve database privacy one has to play a game of balancing two sets of functions: (i) the “private” functions that we wish to hide and (ii) the “information” functions whose values we wish to reveal. This general view allows for a great variety of privacy definitions.

We present a computational definition of privacy that asserts that it is computationally infeasible to retrieve private information from the database. We prefer that to other ‘natural’ measures that were in use in previous works – such as the variance of query answers, and the estimator variance.

There are two potential drawbacks to these definitions. Firstly, it is not clear that large variance necessarily prevents private information from being leaked2. Secondly, this kind of definition does not allow us to capitalize on the limits of an adversary.

One difficulty in estimating (partial) compromise stems from the unknown extent of the adversary’s a-priori knowledge. A way to model prior knowledge is by having the database drawn from some distribution DB over binary strings {0, 1}n. Having no prior knowledge is conceptually equivalent to having all possible database configurations (n-bit strings) equally likely, a situation that is modeled by letting

DB be the uniform distribution over {0, 1}n. Privacy and Cryptography. Privacy is treated in various aspects of cryptography, usually in a manner that is complementary to our discussion. For example, in secure function evaluation several parties compute a function F of their private inputs d1, .., dn. Privacy is perceived here as protecting each party’s private input so that other parties can not deduce information that is not already deducible from the function outcome. In other words, the function F dictates which information is to be revealed, and the goal is to leak no additional information. Note that privacy is defined here implicitly – according to the computed function F, this may lead to leaking no information about the private inputs on one end of the spectrum, and leaking complete information on the other end. In this work we reverse the order. We first specify explicitly which information should not be leaked, and then look for functions revealing the maximum information still possible.

Our privacy vs. information game can be viewed as an interplay between the “private” functions and the “information” functions whose values we wish to compute (or approximate) while maintaining privacy.

IV. MULTIDIMENSIONAL PRIVACY

We now define our privacy criterion. To incorporate external knowledge, the data owner needs to specify the knowledge that an adversary may have. Because it is nearly impossible for the data owner to anticipate the specific knowledge available to an adversary, we take the approach of, and propose a new mechanism for "quantifying" external knowledge. In this approach, the privacy criterion must guarantee safety when the adversary has up to a certain "amount" of knowledge, regardless of the specific things that are known.

As discussed in Section 2.3, in general, it is NP-hard to check safety of a release candidate. Thus, our goal is to find special cases that are both useful in practice and efficiently solvable.

In the rest of this section, we propose an intuitive and usable approach to quantifying adversarial knowledge. The key idea is to break down quantification into several meaningful components, rather than a single number as in. We then define a skyline privacy criterion and a skyline exploratory tool.

V.Three-Dimensional Knowledge

Consider an adversary who wants to determine whether **target individual** *t* (a variable) has **target sensitive value** s (a specific value, e.g., AIDS). Note that *t* is a variable because the target can be anyone, while s is not because we want to provide a possibly different safety guarantee for each unique sensitive value s. Intuitively, we consider the following three types of knowledge:

- *K*s|*t*: Knowledge about the target individual *t*.
- *K*s|*u*: Knowledge about individuals (*u*1, …, *uk*) other than *t*.
- *K*s|*v*,*t*: Knowledge about the relationship between *t* and other individuals (*v*1, …, *vm*).

We note that knowledge about relationships is the most interesting type of knowledge. In this paper, we focus on same-value families, which we consider to be the most natural form of relationship in attribute disclosure. In general, relationships may be expressed using graphs, which is future work.

We use the following convention throughout the paper.

- s is the target sensitive value (a specific value, not a variable).
- *t* is the target individual (a variable).
- *ui*, *vi* are variables ranging over individuals.
- *xi*, *yi* are variables ranging over sensitive values.
- *f*, *g* are (the indices of) QI-groups.

Because the SVPI case and MVPI case have very different characteristics, we discuss these two cases separately.

Calibrating Noise to Sensitivity in Private Data Analysis

1 Introduction

We continue a line of research initiated in [10, 11] on privacy in statistical databases.

A statistic is a quantity computed from a sample. Intuitively, if the database is a representative sample of an underlying population, the goal of a privacy-preserving statistical database is to enable the user to learn properties of the population as a whole while protecting the privacy of the individual contributors.

We assume the database is held by a trusted server. On input a query function

f mapping databases to reels, the so-called true answer is the result of applying f to the database. To protect privacy, the true answer is perturbed by the addition of random noise generated according to a carefully chosen distribution, and this response, the true answer plus noise, is returned to the user. Previous work focused on the case of noisy sums, in which f =Pi g(xi), where xi denotes the it row of the database and g maps database rows to . The power of the noisy sums primitive has been amply demonstrated in, in which it is shown how to carry out many standard data mining and learning tasks using few noisy sum queries.

In this paper we consider general functions f mapping the database to vectors of reals. We prove that privacy can be preserved by calibrating the standard deviation of the noise according to the sensitivity of the function f. This is the maximum amount, over the domain of f, that any single argument to f, that is, any single row in the database, can change the output.

We begin by defining a new notion of privacy leakage. An interaction between a user and a privacy mechanism results in a transcript. For now it is sufficient to think of transcripts corresponding to a single query function and response, but the notion is completely general and our results will handle longer transcripts. Roughly speaking, a privacy mechanism is _-indistinguishable if for all transcripts t and for all databases x and x0 differing in a single row, the probability of obtaining transcript t when the database is x is within a (1 + _) multiplicative factor of the probability of obtaining transcript t when the database is x0. More precisely, we require the absolute value of the logarithm of the ratios to be bounded by _. In our work, _ is a parameter chosen by policy. We then formally define the sensitivity S(f) of a function f. This is a quantity inherent in f; it is not chosen by policy. Note that S(f) is independent of the actual database. The extension to privacy-preserving approximations to "holistic" functions f that operate on the entire database broadens the scope of private data analysis beyond the original motivation of a purely statistical, or "sample population" context. Now we can view the database as an object that is itself of intrinsic interest and that we wish to analyze in a privacy-preserving fashion. For example, the database may describe a concrete interconnection network – not a sample sub network – and we wish to learn certain properties of the network without releasing information about individual edges, nodes, or sub networks. The technology developed herein therefore extends the scope of the line of research, beyond privacy-preserving statistical databases to privacy-preserving analysis of data.

VI.CONCLUSIONS

Microdata privacy can be understood as prevention of membership disclosure (the adversary should not learn whether a particular individual is included in the database) or sensitive attribute disclosure (the sanitized database should not reveal very much information about any individual's sensitive attributes). It is known that generalization and suppression cannot prevent membership disclosure. For sensitive attribute disclosure, perfect privacy can be achieved |against a very weak adversary who knows just the quasi-identifiers by simply removing the sensitive attributes or the quasi-identifiers from the published data. Of course, these trivial sanitizations also destroy any utility that depended on the removed attributes. Algorithms such as k-anonymity and `-diversity leave all sensitive attributes intact and apply generalization and sup- pression to the quasi-identifiers. The goal is to keep the data \truthful" and thus provide good utility for data-mining applications, while achieving less than perfect privacy. We argue that utility is best measured by the success of data mining algorithms such as decision tree learning which take advantage of relationships between attributes. Algorithms that need only aggregate statistical information can be executed on perturbed or randomized data, with much stronger privacy guarantees against stronger adversaries than achieved by k-anonymity, `-diversity, and so on. Our experiments, carried out on the same UCI data as was used to validate existing microdata sanitization algorithms, show that the privacy vs. utility tradeoff for these algorithms is very poor. Depending on the sanitization parameter, sanitized datasets either provide no additional utility vs. trivial sanitization, or the adversary's ability to compute the sensitive attributes of any individual increases much more than the accuracy of legitimate machine-learning workloads. An important question for future research is whether there exists any real-world dataset on which quasi-identifier generalization supports meaningfully better data-mining accuracy than trivial sanitization without severely compromising privacy via sensitive attribute disclosure. Another important question is how to design microdata sanitization algorithms that provide both privacy and utility. Sensitive attribute disclosure results, in part, from the fact that each individual t can only belong to a unique quasi identifier equivalence class in the sanitized table T0. This is a consequence of the requirement that the generalization hierarchy be totally ordered [10]. This requirement helps the adversary, but does not improve utility. If we consider G(t), the set of records in T0 whose quasi-identifier values are generalizations of t[Q], there is no privacy reason why each record of G(t) must have the same quasi-identifier values. It is possible that a generalization strategy that uses, e.g., DAGs instead of totally ordered hierarchies may provide better privacy than the existing algorithms.

VII. REFERENCES

[1] D. N. A. Asuncion. UCI machine learning repository, 2007.

[2] N. Adam and J. Worthmann. Security-control methods for statistical databases: A comparative study. ACM Computing Surveys, 21(4), 1989.

[3] C. Aggarwal. On k-anonymity and the curse of dimensionality. In VLDB, 2005.

[4] R. Agrawal and R. Srikant. Privacy-preserving data mining. In SIGMOD, 2000.

[5] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In ICDE, 2005.

[6] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In PODS, 2005.

[7] J.-W. Byun, Y. Sohn, E. Bertino, and N. Li. Secure anonymization for incremental datasets. In SDM, 2006.

[8] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Towards privacy in public databases. In TCC, 2005.

[9] B.-C. Chen, K. LeFevre, and R. Ramakrishnan. Privacy skyline: privacy with multidimensional adversarial knowledge. In VLDB, 2007.

[10] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. k-anonymity. Secure Data Management in Decentralized Systems, 2007.

[11] I. Dinur and K. Nissim. Revealing information while preserving privacy. In PODS, 2003.

[12] C. Dwork. Di_erential privacy. In ICALP, 2006.

[13] A. Ev_mievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy-preserving data mining. In PODS, 2003.

[14] B. Fung, K. Wang, and P. Yu. Top-down specialization for information and privacy preservation. In ICDE, 2005.

R.Kathyayani.M.Tech,
SRI KRISHNA DEVARAYA ENGINEERING COLLEGE,
JNTU Ananthapuram.

S.G.Nawaz M.Tech, Associate Professor, SRI KRISHNA DEVARAYA ENGINEERING COLLEGE,
JNTU, Ananthapuram.