The Third International Conference on Availability, Reliability and Security

A Critique of k-Anonymity and Some of Its Enhancements

(Invited Paper)

Josep Domingo-Ferrer* and Vicenç Torra**
* Rovira i Virgili University
UNESCO Chair in Data Privacy
Dept. of Computer Engineering and Maths
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
E-mail josep.domingo@urv.cat
** IIIA-CSIC, Campus UAB, E-08193 Bellaterra, Catalonia
E-mail vtorra@iiia.csic.es

Abstract

k-Anonymity is a privacy property requiring that all combinations of key attributes in a database be repeated at least for k records. It has been shown that k-anonymity alone does not always ensure privacy. A number of sophistications of k-anonymity have been proposed, like p-sensitive k-anonymity, l-diversity and t-closeness. This paper explores the shortcomings of those properties, none of which turns out to be completely convincing.

1. Introduction

What is meant by database privacy largely depends on the context where this concept is being used. In official statistics, it normally refers to the privacy of the respondents to which the database records correspond (*respondent privacy*). In co-operative market analysis, it is understood as keeping private the databases owned by the various collaborating corporations (*data owner privacy*). In healthcare, both respondent and owner privacy are implicitly required: patients must keep their privacy and the medical records should not be transferred from a hospital to, say, an insurance company. In the context of dynamically queryable databases and, in particular, Internet search engines, the most rapidly growing concern is *user privacy*, that is, the privacy of the queries submitted by users (especially after scandals like the August 2006 disclosure of 658000 queries by the AOL search engine). Thus, what makes the difference is whose privacy is being sought.

Statistical disclosure control (SDC, [3], [15], [6]) was born in the statistical community as a discipline to achieve respondent privacy. Privacy-preserving data mining (PPDM) appeared simultaneously in the database community [1] and the cryptographic community [8] with the aim of offering owner privacy: several database owners wish to compute queries across their databases in a way that only the results of the queries are revealed to each other, not the contents of each other's databases. Finally, private information retrieval (PIR, [2]) originated in the cryptographic community as an attempt to guarantee the privacy of user queries to databases.

It can be seen that the technologies to deal with the above three privacy dimensions (respondent, owner and user) have evolved in a fairly independent way within research communities with surprisingly little interaction. Fortunately, it turns out that some developments are useful for more than one privacy dimension, even if all three dimensions are independent ([5]). Such is the case for k-anonymity and its evolutions, which are useful properties both for respondent and owner privacy. Furthermore, in combination with private information retrieval, k-anonymity and its evolutions can make all three privacy dimensions compatible.

Therefore, assessing the extent to which those privacy properties really achieve privacy is an important objective that will be pursued in this paper. Section 2 is a critical review of k-anonymity. Section 3 deals with p-sensitive k-anonymity. Section 4 deals with l-diversity. Section 5 deals with t-closeness. Conclusions are drawn in Section 6.

2. k-Anonymity and its shortcomings

k-Anonymity is an interesting approach to facing the conflict between information loss and disclosure risk, suggested by Samarati and Sweeney [11], [10],

0-7695-3102-4/08 $25.00 © 2008 IEEE
DOI 10.1109/ARES.2008.97

[12], [13]. To recall the definition of k-anonymity, we need to enumerate the various (non-disjoint) types of attributes that can appear in a microdata set $\mathbf{X}$:

- *Identifiers*. These are attributes that *unambiguously* identify the respondent. Examples are passport number, social security number, full name, etc. Since our objective is to prevent confidential information from being linked to specific respondents, we will assume in what follows that, in a pre-processing step, identifiers in $\mathbf{X}$ have been removed/encrypted.
- *Key attributes*. Borrowing the definition from [4], [10], key attributes are those in $\mathbf{X}$ that, in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in $\mathbf{X}$ refer. Examples are job, address, age, gender, etc. Unlike identifiers, key attributes cannot be removed from $\mathbf{X}$, because any attribute is potentially a key attribute.
- *Confidential outcome attributes*. These are attributes which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.

k-Anonymity can now be defined as follows.

Definition 1 (k-Anonymity): A protected data set is said to satisfy k-anonymity for $k > 1$ if, for each combination of key attributes, at least k records exist in the data set sharing that combination.

If, for a given k, k-anonymity is assumed to be enough protection for respondents, one can concentrate on minimizing information loss with the only constraint that k-anonymity should be satisfied. This is a clean way of solving the tension between data protection and data utility.

k-Anonymity is able to prevent identity disclosure, *i.e.* a record in the k-anonymized data set cannot be mapped back to the corresponding record in the original data set. However, in general, it may fail to protect against attribute disclosure. This is illustrated by the following example.

Example 1: Imagine that an individual's health record is k-anonymized into a group of k patients with k-anonymized key attributes values Age = "30"*,* Height = "180 cm" *and* Weight = "80 kg"*. Now, if all k patients share the confidential attribute value* Disease = "AIDS"*, k-anonymization is useless, because an intruder who uses the key attributes* (Age, Height, Weight) *can link an external identified record*

(Name="John Smith", Age="31", Height="179",

Weight="81")

with the above group of k patients and infer that John Smith suffers from AIDS (attribute disclosure). □

3. p-Sensitive k-anonymity and its shortcomings

In [14], an evolution of k-anonymity called p-sensitive k-anonymity was presented. Its purpose is to protect against attribute disclosure by requiring that there be at least p different values for each confidential attribute within the records sharing a combination of key attributes. The formal definition is as follows.

Definition 2 (p-Sensitive k-anonymity): A data set is said to satisfy p-sensitive k-anonymity for $k > 1$ and $p \leq k$ if it satisfies k-anonymity and, for each group of records with the same combination of key attribute values the number of distinct values for each confidential attribute is at least p (within the same group).

p-Sensitive k-anonymity has the limitation of implicitly assuming that each confidential attribute takes values uniformly over its domain, that is, that the frequencies of the various values of a confidential attribute are similar. When this is not the case, achieving p-sensitive k-anonymity may cause a huge data utility loss. This is illustrated in the following example.

Example 2: Consider a dataset containing data for 1000 patients. The key attributes in the data set are Age*,* Height *and* Weight*. There is a single confidential attribute* AIDS *whose values can be "Yes" or "No". Assume that there are only five patients in the dataset with* AIDS="Yes"*. Imagine that* 2*-sensitive k-anonymity is desired. Clearly, at least one patient with AIDS is needed in each group sharing a combination of key attributes, so that at most five groups can be formed. Therefore, key attributes must be heavily coarsened so that only five combinations of their values subsist.* □

4. l-Diversity and its shortcomings

Like p-sensitive k-anonymity, l-diversity [9] attempts to solve the attribute disclosure problem that can happen with k-anonymity. We next recall the definion of l-diversity.

Definition 3 (l-Diversity): A data set is said to satisfy l-diversity if, for each group of records sharing a combination of key attributes, there are at least l "well-represented" values for each confidential attribute.

According to [9] the term "well-represented" can be defined in several ways:

1) *Distinct l-diversity*. There must be at least l distinct values for the confidential attribute in

each group of records sharing a combination of key attributes. This is equivalent to l-sensitive k-anonymity and has the same shortcomings.

2) *Entropy l-diversity*. The entropy of a group G for a particular confidential attribute with domain C can be defined as

$$H(G) = -\sum_{c \in C} p(G, c) \log p(G, c)$$

in which $p(G, c)$ is the fraction of records in G which have value c for the sensitive attribute. A dataset is said to have entropy l-diversity if for each group G, $H(G) \geq \log l$.

3) *Recursive (c, l)-diversity*. This property makes sure that the most frequent values do not appear too frequently and the least frequent values do not appear too rarely. Let m be the number of values of the confidential attribute in a group G and r_i, for $1 \leq i \leq m$, be the number of times that the i-th most frequest value appears in G. Then G is said to have recursive (c, l)-diversity if $r_1 < c(r_l + r_{l+1} + \cdots + r_m)$. A dataset is said to have recursive (c, l)-diversity if all its groups have recursive (c, l)-diversity. A special case of recursive (c, l)-diversity is (α, k)-anonymity [17], which requires that the proportion of each sensitive value in each group is at most $\alpha \in [0, 1]$.

In [7], the following criticisms are made to l-diversity:

- *l-Diversity may be difficult and unnecessary to achieve*. The argument is the same given against p-sensitive k-anonymity in Example 2 above.
- *l-Diversity is insufficient to prevent attribute disclosure*. At least the following two attacks are conceivable:
  - *Skewness attack*. If, in Example 2, a group has the same number of patients with and without AIDS; in that case, it satisfies distinct 2-diversity, entropy 2-diversity and any recursive $(c, 2)$-diversity requirement. However, if an intruder can link a specific patient to that group, that patient can be considered to have 50% probability of having AIDS, in front of $5/1000$ for the overall data set.
  - *Similarity attack*. If values of a sensitive attribute within a group are l-diverse but semantically similar, attribute disclosure also takes place. *E.g.* if patients in a 3-diverse data set where *Disease* is a confidential attribute all have values in

    {"lung cancer", "liver cancer", "stomach cancer"}

    an intruder linking a specific individual to that group can infer that the individual has cancer. If the confidential attribute is numerical and values within a group are l-diverse but very similar, the intruder can estimate the confidential attribute value for an individual in that group to a narrow interval.

5. t-Closeness and its shortcomings

In [7], a new privacy property called t-closeness is defined as follows.

Definition 4 (t-Closeness): A data set is said to satisfy t-closeness if, for each group of records sharing a combination of key attributes, the distance between the distribution of the confidential attribute in the group and the distribution of the attribute in the whole data set is no more than a threshold t.

t-Closeness solves the attribute disclosure vulnerabilities inherent to l-diversity:

- *Skewness attack*. Since the within-group distribution of confidential attributes is the same as the distribution of those attributes for the entire dataset, no skewness attack can occur.
- *Similarity attack*. Again, since the within-group distribution of confidential attributes mimics the distribution of those attributes over the entire dataset, no semantic similarity can occur within a group that does not occur in the entire dataset. (Of course, within-group similarity cannot be avoided if all patients in a data set have similar diseases.)

However, some criticisms can be made to t-closeness:

- Whereas the paper [7] elaborates on several ways to check t-closeness (using several distances between distributions), no computational procedure to enforce this property is given.
- If such a procedure was available, it would greatly damage the utility of data. Indeed, the authors of [7] acknowledge that t-closeness limits the amount of useful information that is released. In fact, limiting is too mild a word, because enforcing t-closeness destroys the correlations between key attributes and confidential attributes: by definition of t-closeness the values of a confidential attribute have the same distribution for any combination of values of key attributes! The only way to decrease the damage is to increase the threshold t, that is, to relax t-closeness.

6. Conclusions and future research

Neither k-anonymity nor its enhancements examined in this paper are entirely successful in ensuring that no privacy leakage occurs while keeping a reasonable data utility level. In fact, while k-anonymity, p-sensitive k-anonymity and l-diversity do not completely protect privacy, t-closeness offers complete privacy at the cost of severely impairing the correlations between confidential attributes and key attributes.

Another problem of the above properties is the computational approach to reach them for a specific dataset to be anonymized. The papers defining k-anonymity, p-sensitive k-anonymity and l-diversity propose approaches based on generalization and suppression which, among other shortcomings, fail to preserve the nature of numerical attributes by causing them to become categorical. In the case of t-closeness, there is not even mention of a computational procedure to reach it.

Therefore, there are plenty of open research avenues in this area, both at the conceptual level (definition of better properties) and at the computational level (definition of less disruptive computational procedures). If, in addition, one assumes that the intruder knows the precise privacy property being pursued by the data protector (as assumed in [16]), new challenges appear.

Disclaimer and acknowledgments

The authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization. This work was partly supported by the Spanish Ministry of Education through projects TSI2007-65406-C03-01 "E-AEGIS" and CONSOLIDER CSD2007-00004 "ARES", and by the Government of Catalonia under grant 2005 SGR 00446.

References

[1] R. Agrawal and R. Srikant. Privacy preserving data mining. In *Proceedings of the ACM SIGMOD*, pages 439–450. ACM, 2000.

[2] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan. Private information retrieval. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 41–50, 1995.

[3] T. Dalenius. The invasion of privacy problem and statistics production. an overview. *Statistik Tidskrift*, 12:213–225, 1974.

[4] T. Dalenius. Finding a needle in a haystack - or identifying anonymous census records. *Journal of Official Statistics*, 2(3):329–336, 1986.

[5] J. Domingo-Ferrer. A three-dimensional conceptual framework for database privacy. In *Secure Data Management-4th VLDB Workshop SDM'2007*, volume 4721 of *Lecture Notes in Computer Science*, pages 193–202, Berlin Heidelberg, 2007.

[6] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte-Nordholt, G. Seri, and P.-P. DeWolf. *Handbook on Statistical Disclosure Control (version 1.0)*. Eurostat (CENEX SDC Project Deliverable), 2006.

[7] N. Li, T. Li, and S. Venkatasubramanian. T-closeness: privacy beyond k-anonymity and l-diversity. In *Proceedings of the IEEE ICDE 2007*, 2007.

[8] Y. Lindell and B. Pinkas. Privacy preserving data mining. In *Advances in Cryptology - CRYPTO'00*, volume 1880 of *Lecture Notes in Computer Science*, pages 36–53, Berlin Heidelberg, 2000.

[9] A. Machanavajjhala, J. Gehrke, D. Kiefer, and M. Venkitasubramanian. L-diversity: privacy beyond k-anonymity. In *Proceedings of the IEEE ICDE 2006*, 2006.

[10] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[11] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International, 1998.

[12] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):571–588, 2002.

[13] L. Sweeney. k-anonimity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):557–570, 2002.

[14] T. M. Truta and B. Vinay. Privacy protection: p-sensitive k-anonymity property. In *2nd International Workshop on Privacy Data Management PDM 2006*, page p. 94, Berlin Heidelberg, 2006. IEEE Computer Society.

[15] L. Willenborg and T. DeWaal. *Elements of Statistical Disclosure Control*. Springer-Verlag, New York, 2001.

[16] R. C.-W. Wong, A. W.-C. Fu, Ke Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In *Proceedings of the VLDB 2007*, Vienna, 2007.

[17] R. C.-W. Wong, J. Li, A. W.-C. Fu, and Ke Wang. (α,k)-anonymity: an enhanced k-anonymity model for privacy-preserving data publishing. In *Proceedings of the ACM KDD*, pages 754–759, New York, 2006.