A Three-Dimensional Conceptual Framework for Database Privacy

Josep Domingo-Ferrer

Rovira i Virgili University
UNESCO Chair in Data Privacy
Department of Computer Engineering and Mathematics
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
josep.domingo@urv.cat

Abstract. Database privacy is an ambiguous concept, whose meaning is usually context-dependent. We give a conceptual framework for technologies in that field in terms of three dimensions, depending on whose privacy is considered: i) respondent privacy (to avoid re-identification of patients or other individuals to whom the database records refer); ii) owner privacy (to ensure that the owner must not give away his dataset); and iii) user privacy (to preserve the privacy of queries submitted by a data user). Examples are given to clarify why these are three *independent* dimensions. Some of the pitfalls related to combining the privacy interests of respondents, owners and users are discussed. An assessment of database privacy technologies against the three dimensions is also included.

Keywords: Statistical database privacy, Private information retrieval, Privacy-preserving data mining, Security and privacy of electronic health records.

1 Introduction

The meaning of database privacy is largely dependent 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. In co-operative market analysis, it is understood as keeping private the databases owned by the various collaborating corporations. In healthcare, both of the above requirements may be implicit: 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 interactively queryable databases and, in particular, Internet search engines, the most rapidly growing concern is the privacy of the queries submitted by users (especially after scandals like the August 2006 disclosure by the AOL search engine of 36 million queries made by 657000 users). Thus, what makes the difference is whose privacy is being sought.

The last remark motivates splitting database privacy in the following three dimensions:

W. Jonker and M. Petković (Eds.): SDM 2007, LNCS 4721, pp. 193–202, 2007.

1. *Respondent privacy* is about preventing re-identification of the respondents (*e.g.* individuals like patients or organizations like enterprises) to which the records of a database correspond. Usually respondent privacy becomes an issue only when the database is to be made available by the data collector (hospital or national statistical office) to third parties, like researchers or the public at large.
2. *Owner privacy* is about two or more autonomous entities being able to compute queries across their databases in such a way that only the results of the query are revealed.
3. *User privacy* is about guaranteeing the privacy of queries to interactive databases, in order to prevent user profiling and re-identification.

The technologies to deal with the above three privacy dimensions have evolved in a fairly independent way within research communities with surprisingly little interaction:

- Respondent privacy is pursued mainly by statisticians and a few computer scientists working in statistical disclosure control (SDC), also known as statistical disclosure limitation (SDL) or inference control [17,26].
- Owner privacy is the primary though not the only goal [1] of privacy-preserving data mining (PPDM, [5]), a discipline born in the database and data mining community. Interestingly enough, the term privacy-preserving data mining independently and simultaneously appeared in the cryptographic community [18,19] to denote a special case of secure multiparty computation where each party holds a subset of the records in a database (horizontal partitioning).
- Finally, user privacy has found solutions mainly in the cryptographic community, where the notion of private information retrieval was invented (PIR, [8]).

Only quite recently some researchers have started to look for holistic privacy solutions, but to the best of our knowledge no comprehensive technology covering the three dimensions above exists yet. In [6] the apparent conflict between respondent privacy and user privacy is highlighted: it seems necessary for the data owner to analyze the user queries in order to check that they will not result in disclosure of sensitive respondent data. In [3], a new technology called hippocratic databases is described which seeks to ensure respondent privacy and owner privacy; examples related to healthcare are provided, with respondents being patients and data owners being hospitals.

1.1 Contribution and Plan of this Paper

The general aim of this paper is to clarify the independent nature of the privacy of respondents, owners and users in databases. Specifically, it will be shown

[1] In [13], a PPDM approach based on randomized responses is presented whose primary aim is claimed to be respondent privacy; however, typical respondents are unlikely to have or use the proposed randomizing device when answering a survey, whereas data owners could make use of it to protect their own privacy.

that guaranteeing privacy for one of those entities does not ensure privacy for the other two. These three privacy dimensions constitute a conceptual framework that can be used to classify technologies based on whose privacy they offer.

In Section 2 we show that respondent privacy and owner privacy are independent dimensions. Section 3 illustrates the independence of respondent privacy and user privacy. Section 4 deals with the independence of owner privacy and user privacy. A tentative assessment of technologies according to the three privacy dimensions is given in Section 5. Section 6 contains some conclusions on the simultaneous satisfaction of the three privacy dimensions as well as topics for future research.

2 Independence of Respondent Privacy vs Owner Privacy

If a dataset is published without any anonymization masking, in general it will violate both respondent and owner privacy. However, there are more interesting cases.

Respondent privacy without owner privacy. Consider the patient toy dataset 1 in Table 1 (left). Assume that the records have been obtained by a pharmaceutical company which is testing a new drug against hypertension. All patients in the dataset suffered from hypertension before starting the treatment. Direct identifiers have been suppressed, but height and weight constitute key attributes in the sense of [9,20], *i.e.* they identify the respondent with some degree of ambiguity [2]: an intruder can easily gauge the height and weight of an individual he knows in order to link the identity of that individual to a record in the dataset. The remaining attributes (systolic blood pressure and AIDS) are confidential attributes.

Luckily enough for the patients in Dataset 1 of Table 1, the dataset turns out to spontaneously satisfy k-anonymity [21,20,23] for $k = 3$ with respect to the key attributes (height, weight). In other words, each combination of the key attributes appears at least 3 times. Thus, if 3-anonymity is considered to be enough protection for patients, Dataset 1 offers respondent privacy and could be published as far as the patients are concerned [3].

However, since Dataset 1 is the one actually obtained in the clinical drug trial, the pharmaceutical company is unwilling to share those data with possible

[2] Height and weight make sense as key attributes only in small populations. In large populations, there are many individuals sharing similar weights and heights. However, we keep those two attributes as example key attributes for the sake of clarity.

[3] If records sharing a combination of key attributes in a k-anonymous dataset also share the values for one or more confidential attributes, then k-anonymity does not guarantee respondent privacy. A stronger property called p-sensitive k-anonymity [24] is in general required: there should be at least p distinct values of each confidential attribute within each group of records sharing a combination of key attributes.

Table 1. Left, patient data set no. 1. Right, patient data set no. 2.

Height (cm)	Weight (kg)	Blood pressure (syst, mmHg)	AIDS (Y/N)
175	76	117	Y
175	76	131	N
175	76	122	N
180	81	115	N
180	81	122	Y
180	81	146	N
190	95	110	N
190	95	115	Y
190	95	125	N
190	95	140	N

Height (cm)	Weight (kg)	Blood pressure (syst, mmHg)	AIDS (Y/N)
160	110	146	N
170	65	117	Y
173	75	131	N
175	80	122	N
180	68	115	N
183	81	122	Y
187	95	110	N
190	95	115	Y
192	99	125	N
192	101	140	N

competitors. Therefore publication of the dataset is compatible with respondent privacy but violates the owner privacy.

Respondent privacy and owner privacy. If a dataset is adequately masked before release, then both owner and respondent privacy are obtained without significantly damaging the utility of the data for designated user analyses. There are plenty of examples in the literature along this line, some of which are:

- In [5], noise addition is used to mask an original dataset for owner privacy and, to a large extent, for respondent privacy. Regarding utility, the distribution of the original dataset can still be reconstructed from the noise-added data, so that decision-tree classifiers properly run on the masked data.
- In [1], masking through condensation (actually a special case of multivariate microaggregation, [10]) is proposed to achieve privacy-preserving data mining. Since the covariance structure of the original attributes is preserved, a variety of analyses can be validly carried out by users on the masked data. Since microaggregation/condensation with minimum group size k on the key attributes guarantees k-anonymity ([12]), the approach in [1] can also guarantee respondent privacy.

Hippocratic databases [4,3] mentioned above are a real-world technology integrating k-anonymization for respondent privacy and PPDM based on noise addition [15] for owner privacy.

Owner privacy without respondent privacy. Imagine that the patient dataset obtained by the pharmaceutical company is not Dataset 1 in Table 1, but Dataset 2. The new dataset is no longer 3-anonymous with respect to the key attributes (height, weight). Therefore, releasing a single record is a violation of respondent privacy: the patient's blood pressure and AIDS condition could be linked to his/her identity by means of the unique patient's key attributes. In fact, merely revealing the name of a patient who took part in the trial already discloses that he/she suffers from hypertension (only patients with hypertension underwent the trial).

However, neither revealing a single record nor the name of someone who took part in the trial can be said to violate the data owner's privacy (especially if the dataset is large). Thus, we can have owner privacy without respondent privacy.

A subtler example is conceivable with the method proposed in [5] and mentioned above. In [11] it is shown that, for higher-dimensional datasets, the property of the method in [5] that the distribution of the original data can be reconstructed from the noise-added data can result in violation of respondent privacy. The reason is that, for higher dimensions, data tend to become sparse, *i.e.* with a lot of rare combinations of attribute values: if the reconstructed distribution fits the multidimensional histogram of the original data too well, rare combinations in the original data are disclosed. This is a non-trivial case of owner privacy without respondent privacy.

3 Independence of Respondent Privacy and User Privacy

The trivial case with neither respondent nor user privacy is the most common one: a queryable database where neither records nor user queries undergo any anonymization (in particular, this is the case of Internet search engines). Situations with at least respondent or user privacy are discussed in the next subsections.

Respondent privacy without user privacy. The conflict between respondent privacy and user privacy is apparent in statistical disclosure control of interactively queryable statistical databases. The scenario is a database to which the user can submit statistical queries (sums, averages, etc.). The aggregate information obtained by a user as a result of successive queries should not allow him to infer the values of confidential attributes for specific individuals (respondent privacy). Currently employed strategies rely on perturbing, restricting or replacing by intervals the answers to certain queries. Examples of those three strategies can be found in [7,14,16], respectively.

All SDC methods for interactive statistical databases assume that the data owner operating the database exactly knows the queries submitted by users. This knowledge is deemed necessary to check that users do not submit a series of queries designed to isolate a single record in the database. Thus, there is no user privacy whatsoever. Even without user privacy, the SDC problem in this kind of databases is known to be difficult since the 1980s, due to the existence of the tracker attack [22].

Respondent privacy and user privacy. If the records in an interactively queryable statistical database are k-anonymous (spontaneously as in Dataset 1 or after a k-anonymization process as described in [2,12]), then no user query can jeopardize respondent privacy. In this case, the use of private information retrieval protocols to preserve the privacy of user queries can be afforded.

User privacy without respondent privacy. This situation is the most likely one if private information retrieval is allowed on unmasked records. To illustrate, assume that PIR is offered on Dataset 2 in Table 1. Even if allowed queries

are only of statistical nature, a user could take advantage of PIR to submit the following queries (assuming PIR protocols existed for those query types):

```
SELECT COUNT(*) FROM Dataset 2 WHERE height < 165 AND weight > 105
SELECT AVG(blood_pressure) FROM Dataset 2 WHERE height < 165 AND
weight > 105
```

The first query tells the user that there is only one individual in the dataset smaller than 165 cm and heavier than 105 kg. With this knowledge, the user can establish that the average blood pressure 146 returned by the second query corresponds to that single individual, who turns out to be someone suffering from serious hypertension. Re-identifying such a small and heavy individual as Mr./Mrs. X should not be too difficult. If the user is an insurance company, Mr./Mrs. X might see his/her life insurance application rejected or accepted only at an extremely high premium.

4 Independence of Owner Privacy and User Privacy

If a database owner allows unrestricted queries on original data and user queries are not protected, there is neither owner privacy nor user privacy. The cases with at least one of both properties are next discussed.

Owner privacy without user privacy. PPDM methods developed in the cryptographic community in the spirit of the seminal paper [18] are special cases of secure multiparty computation. The idea is that two or more parties owning confidential databases run a data mining algorithm (*e.g.* a classifier) on the union of their databases, without revealing any unnecessary information.

Thus, the users coincide with the data owners. However, the focus is on ensuring the privacy of the data owned by each party (owner privacy). The analysis or data mining algorithm run by the parties is known to all of them. Indeed, as usual in secure multiparty computation, all parties interactively co-operate to obtain the result of the analysis. This is hardly compatible with private information retrieval for user privacy. Thus, we can conclude that cryptographic PPDM offers owner privacy but no user privacy.

Non-cryptographic PPDM methods developed in the data mining community are friendlier toward user privacy, as will be discussed in the next paragraph below. However, application of those methods without PIR also leads to owner privacy without user privacy.

Owner privacy and user privacy. Unlike cryptographic PPDM, non-cryptographic PPDM developed by data miners is usually non-interactive. The data owner first protects his data and then accepts queries on them. Whatever the case, the data owner does not need to know the exact query being computed on his protected data, so that PIR for user privacy is compatible with non-cryptographic PPDM.

One must acknowledge here that, while some PPDM methods like [2] allow a broad range of analyses/queries to be performed on the protected data, other

methods have been designed to support a specific class of analyses on the privacy-protected data (*e.g.* [5] is designed for decision-tree classifiers and methods in [25] are designed for association rule mining). However, even with the latter methods, the data owner does not need to know anything about the exact user analyses beyond the (likely) fact that they belong to the class supported by the PPDM algorithm.

User privacy without owner privacy. This is the situation if unrestricted PIR queries are allowed by an owner on his original data. If the database is a public one and contains non-confidential information, this is the most desirable situation. For example, in the context of Internet search engines, user privacy is arguably the only privacy that should be cared about.

5 Tentative Technology Scoring

In order to demonstrate the usefulness of the proposed three-dimensional conceptual framework, we attempt as an exercise a scoring of the *non-exhaustive* list of privacy technologies mentioned in this paper. Table 2 is an assessment of how well each technology class performs in each privacy dimension. This scoring is qualitative and tentative, in that we base our assessment on the usual claims of each technology class, rather than on the actual properties of specific proposals within each class. For the reader's orientation, we mean by SDC the methods in [17,26]; example proposals of use-specific non-crypto PPDM are [5,25]; an example generic non-crypto PPDM method is [2]; an example PIR method is [8].

Table 2. Tentative scoring of technology classes

Technology class	Respondent privacy	Owner privacy	User privacy
SDC	medium-high	medium	none
Use-specific non-crypto PPDM	medium	medium-high	none
Generic non-crypto PPDM	medium	medium-high	none
Crypto PPDM	high	high	none
PIR	none	none	high
SDC + PIR	medium-high	medium	high
Use-specific non-crypto PPDM + PIR	medium	medium-high	medium
Generic non-crypto PPDM + PIR	medium	medium-high	high

The rationale for the grades in Table 2 follows:

– Being based on multi-party secure computation, crypto PPDM methods are the PPDM methods offering highest owner privacy. As a side property, they also offer respondent privacy (records in the database are not leaked). In comparison, non-crypto PPDM only offers medium-high owner privacy; however, as argued above, it is more flexible and it can be combined with PIR.

- A distinction is made between use-specific and generic non-crypto PPDM: when use-specific non-crypto PPDM is combined with PIR, there is some clue on the queries made by the user (they are likely to correspond to the uses the PPDM method is intended for); therefore generic non-crypto PPDM is better for combination with PIR in view of attaining high user privacy.
- Non-crypto PPDM and SDC in Table 2 are assumed to rely on data masking, rather than on query control.
- If non-crypto PPDM perturbs the data, it normally provides some level of respondent privacy in addition to owner privacy.
- Similarly, SDC masking normally provides some level of owner privacy in addition to respondent privacy.

Note that the last two assertions do not contradict the independence between respondent and owner privacy, justified in Section 2 above.

6 Conclusions and Future Research

Respondent privacy, owner privacy and user privacy have been shown to be independent, yet compatible properties. Even though satisfying one of them gives no assurance about the others, we can state a few lessons learned which can be used as guidelines to simultaneous fulfillment of the three privacy dimensions:

- Respondent privacy relies on data masking (*e.g.* for k-anonymity) or on query control (needed if interactive queries against original databases are allowed). Since query control is hardly compatible with user privacy, data masking must be used for respondent privacy if the latter property must live together with user privacy.
- Owner privacy relies on cryptographic or non-cryptographic PPDM. Being based on interactive multiparty computation, cryptographic PPDM assumes that the joint computation being carried out is known to all parties, which is not compatible with user privacy. Therefore, non-cryptographic PPDM seems a wiser choice if owner privacy is to be made compatible with user privacy.
- Most forms of non-cryptographic PPDM rely on perturbing the original data. If this perturbation is such that the underlying data are k-anonymized (as in [2,12]), then owner and respondent privacy are simultaneously achieved.

Hence, one possible way to fulfill the three privacy dimensions is for a database which is not originally k-anonymous to be k-anonymized (via microaggregation-condensation, recoding, suppression, etc.) and to be added a PIR protocol to protect user queries.

Future research should explore other possible solutions satisfying the privacy of respondents, owners and users. Also, the impact on data utility of offering the three dimensions of privacy (rather than just one or two of them) should be investigated. An interesting challenge is to offer privacy for everyone without incurring extra data utility penalties.

Disclaimer and Acknowledgments

The author is 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 SEG2004-04352-C04-01 "PROPRIETAS" and CONSOLIDER CSD2007-00004 "ARES", and by the Government of Catalonia under grant 2005 SGR 00446. Thanks go to Agusti Solanas for his comments on a draft version of this paper.

References

1. Aggarwal, C.C., Yu, P.S.: A condensation approach to privacy preserving data mining. In: Bertino, E., Christodoulakis, S., Plexousakis, D., Christophides, V., Koubarakis, M., Böhm, K., Ferrari, E. (eds.) EDBT 2004. LNCS, vol. 2992, pp. 183–199. Springer, Heidelberg (2004)
2. Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., Zhu, A.: k-Anonymity: Algorithms and hardness. Technical report, Stanford University (2004)
3. Agrawal, R., Grandison, T., Johnson, C., Kiernan, J.: Enabling the 21st century health care information technology revolution. Communications of the ACM 50(2), 35–42 (2007)
4. Agrawal, R., Kiernan, J., Srikant, R., Xu, Y.: Hippocratic databases. In: Proceedings of the 28th International Conference on Very Large Databases, Hong Kong (2002)
5. Agrawal, R., Srikant, R.: Privacy preserving data mining. In: Proceedings of the ACM SIGMOD, pp. 439–450. ACM Press, New York (2000)
6. Aguilar, C., Deswarte, Y.: Single database private information retrieval schemes. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 257–265. Springer, Heidelberg (2006)
7. Chin, F.Y., Ozsoyoglu, G.: Auditing and inference control in statistical databases. IEEE Transactions on Software Engineering E-8, 574–582 (1982)
8. Chor, B., Goldreich, O., Kushilevitz, E., Sudan, M.: Private information retrieval. In: IEEE Symposium on Foundations of Computer Science (FOCS), pp. 41–50. IEEE Computer Society Press, Los Alamitos (1995)
9. Dalenius, T.: Finding a needle in a haystack - or identifying anonymous census records. Journal of Official Statistics 2(3), 329–336 (1986)
10. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering 14(1), 189–201 (2002)
11. Domingo-Ferrer, J., Sebé, F., Castellà, J.: On the security of noise addition for privacy in statistical databases. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 149–161. Springer, Heidelberg (2004)
12. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogenerous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2), 195–212 (2005)
13. Du, W., Zhan, Z.: Using randomized response techniques for privacy-preserving data mining. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C, pp. 505–510 (2003)

14. Duncan, G.T., Mukherjee, S.: Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise. Journal of the American Statistical Association 95, 720–729 (2000)
15. Evfimievski, A.: Randomization in privacy-preserving data mining. SIGKDD Explorations: Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining 4(2), 43–48 (2002)
16. Gopal, R., Garfinkel, R., Goes, P.: Confidentiality via camouflage: the cvc approach to disclosure limitation when answering queries to databases. Operations Research 50, 501–516 (2002)
17. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte-Nordholt, E., Seri, G., DeWolf, P.-P.: Handbook on Statistical Disclosure Control (version 1.0). In: Eurostat (CENEX SDC Project Deliverable) (2006)
18. Lindell, Y., Pinkas, B.: Privacy preserving data mining. In: Bellare, M. (ed.) CRYPTO 2000. LNCS, vol. 1880, pp. 36–53. Springer, Heidelberg (2000)
19. Lindell, Y., Pinkas, B.: Privacy preserving data mining. Journal of Cryptology 15(3), 177–206 (2002)
20. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
21. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International (1998)
22. Schlörer, J.: Disclosure from statistical databases: quantitative aspects of trackers. ACM Transactions on Database Systems 5, 467–492 (1980)
23. Sweeney, L.: k-Anonimity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 10(5), 557–570 (2002)
24. Truta, T.M., Vinay, B.: Privacy protection: p-sensitive k-anonymity property. In: 2nd International Workshop on Privacy Data Management PDM 2006, p. 94. IEEE Computer Society Press, Los Alamitos (2006)
25. Verykios, V.S., Elmagarmid, A.K., Bertino, E., Saygin, Y., Dasseni, E.: Association rule hiding. IEEE Transactions on Knowledge and Data Engineering 16(4), 434–447 (2004)
26. Willenborg, L., DeWaal, T.: Elements of Statistical Disclosure Control. Springer, New York (2001)