A Three-Dimensional Conceptual Framework for Database Privacy

Josep Domingo-Ferrer
Rovira i Virgili University

http://crises-deim.urv.cat/ares

May 15, 2008

Introduction

The meaning of database privacy is context-dependent:

- In official statistics, it normally refers to privacy of the respondents to which database records correspond.
- In co-operative market analysis, it means keeping private the databases owned by the various collaborating organizations.
- In healthcare, both requirements above may be implicit: patients should keep their privacy and the hospital should have control on its medical records.
- In interactive databases and Internet search engines, the privacy of queries submitted by users is a growing concern [1].

[1] Especially after the August 2006 disclosure of queries by 658000 users by AOL!

The three privacy dimensions

Depending on whose privacy it is being sought, database privacy can be split in three dimensions:

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 dynamic databases, in order to prevent user profiling and re-identification.

Technologies for DB privacy

- Respondent privacy is pursued mainly by statisticians and a few computer scientists working in statistical disclosure control (SDC).
- Owner privacy is the goal of privacy-preserving data mining (PPDM), a discipline born in the database and data mining community. Privacy-preserving data mining independently and simultaneously appeared in the cryptographic community 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).

Transversal developments

- The technologies to deal with the three privacy dimensions have evolved fairly independently within communities with surprisingly little interaction.
- No comprehensive technology covering the three dimensions above exists yet.
- In Aguilar and Deswarte (2006), the apparent conflict between respondent privacy and user privacy is highlighted: it seems necessary for the data owner to analyze user queries in order to guarantee respondent privacy.
- Agrawal *et al.* (2007) propose hippocratic databases to ensure both respondent and owner privacy, especially in healthcare applications.

Objectives of this presentation

- Clarify the independent nature of the privacy of respondents, owners and users of databases.
- Show that guaranteeing privacy for one of such entities does not ensure privacy for the other two.
- Assess technologies according to whose privacy they offer.

Outline of contents

- Independence of respondent and owner privacy
- Independence of respondent and user privacy
- Independence of owner privacy and user privacy
- Assessment of privacy technologies
- Conclusion and open issues

Independence of respondent and owner privacy

- If a dataset is published without anonymization masking, in general it violates both respondent and owner privacy.
- But respondent privacy can exist without owner privacy and conversely.

Respondent privacy without owner privacy

Consider a dataset whose records have been obtained by a pharmaceutical company testing a new drug against hypertension:

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

Respondent privacy without owner privacy (II)

- All patients in the dataset suffered from hypertension before starting the treatment.
- Direct identifiers have been suppressed.
- Height and weight are key attributes: an intruder can easily gauge the height and weight of an individual he knows so as to link her identity to a record in the dataset.
- The remaining attributes (systolic blood pressure and AIDS) are confidential attributes.
- The dataset spontaneously satisfies 3-anonymity for the key attributes height and weight.

$\Longrightarrow$ If 3-anonymity is enough protection for patients, release of the dataset results does not harm respondent privacy but it harms 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.
- E.g. Agrawal and Srikant (2000) use noise addition for owner privacy and, as a by-product, respondent privacy.
- E.g. Aggarwal and Yu (2004) use a special kind of multivariate microaggregation for k-anonymization to attain the same purpose.
- Hippocratic databases integrate k-anonymization for respondent privacy and PPDM based on noise addition.

Owner privacy without respondent privacy

Consider the dataset:

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

Owner privacy without respondent privacy (II)

- This dataset is no longer 3-anonymous with respect to key attributes height and weight.
- Releasing a single record is a violation of respondent privacy: the patient's blood pressure and AIDS condition can be linked to her identity and her hypertension is leaked out.
- 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).
- The noise-based PPDM scheme by Agrawal and Srikant (2000) is another example: owner privacy is preserved but the privacy of respondents with rare combinations of attribute values is not guaranteed.

Independence of respondent and user privacy

- The trivial case with neither respondent nor user privacy is the most common.
- This is the case a queryable database where neither records nor user queries undergo any anonymization (*e.g.* an Internet search engine).

Respondent privacy without user privacy

- The conflict between respondent privacy and user privacy is apparent in SDC of interactively queryable statistical databases.
- A user submits statistical queries to the database (sums, averages, etc.).
- Successive queries should not allow the user to infer the values of confidential attributes for specific individuals (respondent privacy).
- Current strategies for this include query perturbation, query restriction and interval answers.
- The database is always assumed to exactly know the queries submitted by users (no user privacy).

Respondent privacy and user privacy

- If the records in an interactively queryable statistical database are k-anonymous (spontaneously or after k-anonymization), then no user query can jeopardize respondent privacy.
- In this case, the use of private information retrieval (PIR) protocols to preserve user privacy can be afforded.

User privacy without respondent privacy

- This situation is the most likely one if PIR is allowed on unmasked records.
- Even if allowed queries are only statistical, consider
  `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.

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.

Owner privacy without user privacy

- Cryptographic PPDM methods are special cases of secure multiparty computation: several parties owning confidential databases want to compute on the union of those databases.
- Thus, the users coincide with the data owners.
- The focus is on owner privacy.
- The kind of computation carried out is known to all parties (no user privacy).

Owner privacy and user privacy

- Non-cryptographic PPDM developed by data miners is usually non-interactive.
- Data are first protected (using noise addition or microaggregation).
- Then queries are accepted on the protected data.
- The data owner does not need to know the exact query being computed on his protected data, so that PIR is compatible with non-cryptographic PPDM.
- Still, some non-cryptographic PPDM methods are designed only for a specific class of analyses/queries on the protected data, which slightly limits user privacy.

User privacy without owner privacy

- This is the situation if unrestricted PIR is allowed by an owner on his original data.
- This is the most desirable situation if the database is public, as it happens in the context of Internet search engines, where only user privacy should matter.

Tentative assessment of privacy technologies

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

Rationale

- Crypto PPDM methods offer highest owner privacy. They also offer respondent privacy (records in the database are not leaked). Non-crypto PPDM only offers medium-high owner privacy; however, it is more flexible and it can be combined with PIR.
- When use-specific non-crypto PPDM is combined with PIR, there is some clue on the queries made by the user; therefore generic non-crypto PPDM is better for combination with PIR in view of attaining high user privacy.
- Non-crypto PPDM and SDC are assumed to rely on data masking, rather than on query control.
- If non-crypto PPDM perturbs the data, it normally provides some respondent privacy in addition to owner privacy.
- Similarly, SDC masking normally provides some owner privacy in addition to respondent privacy.

Conclusions

- Respondent privacy, owner privacy and user privacy have been shown to be independent, yet compatible properties.
- Some guidelines to simultaneous fulfillment of the three privacy dimensions are:
  - Respondent privacy relies on data masking or on query control. Query control is hardly compatible with user privacy, so data masking must be used for respondent privacy to be compatible with user privacy.
  - Owner privacy relies on crypto or non-crypto PPDM. Crypto PPDM assumes that the target computation is known to all parties, so no user privacy $\Longrightarrow$ non-crypto PPDM seems a wiser choice to make owner and user privacy compatible.
  - Most forms of non-crypto PPDM rely on perturbing the original data. If perturbation k-anonymizes the data, then owner and respondent privacy are simultaneously achieved.

Open issues

- 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.
- Other possible solutions satisfying the privacy of respondents, owners and users should be explored.
- 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.