Data & Knowledge Engineering xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Editorial

Recent progress in database privacy

Josep Domingo-Ferrer [a,*], Yücel Saygın [b]

[a] Universitat Rovira i Virgili, Dept. of Computer Engineering and Mathematics, UNESCO Chair in Data Privacy, Av. Països Catalans 26,
E-43007 Tarragona, Catalonia
[b] Sabancı University, Faculty of Engineering and Natural Sciences, Orhanli, 34956, Tuzla, Istanbul, Turkey

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Database privacy
Statistical disclosure control
Privacy-preserving data mining
Private information retrieval

ABSTRACT

Database privacy can be conceptually described in terms of three dimensions, each of which refers to the privacy of a different subject and is pursued by a different discipline: respondent privacy, addressed by statistical disclosure control (SDC), owner privacy, addressed by privacy-preserving data mining (PPDM), and user privacy, pursued by private information retrieval (PIR). This special issue contains papers reporting recent advances in each of those three dimensions.

1. Introduction

Database privacy is a multifaceted property, depending on whose privacy is being sought. In official statistics, it normally refers to the privacy of respondents having supplied the information to which the database records correspond. In e-commerce and industry, those who care most about privacy are often the database owners, for whom data collection has a substantial cost and/or represents a competitive edge; think for example of data collected by a pharmaceutical company in several costly drug trials. In Internet search engines, the most rapidly growing concern is the privacy of the queries submitted by users.

Thus, what makes the difference is whose privacy is being sought. Based on those remarks, in [3] three privacy dimensions were distinguished:

(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 (health or statistical agency) 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 [5,6].

[*] Corresponding author. Tel.: +34 977558109; fax: +34 977559710.
E-mail addresses: josep.domingo@urv.cat (J. Domingo-Ferrer), ysaygin@sabanciuniv.edu (Y. Saygın).

0169-023X/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2009.06.002

- Owner privacy is the primary though not the only goal of privacy-preserving data mining (PPDM [1]), 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 [7] to denote a special case of secure multiparty computation where each party holds a subset of the records in a database.
- Finally, user privacy has found solutions mainly in the cryptographic community, where the notion of private information retrieval was invented (PIR [2]).

One ambition of this special issue is to offer an opportunity for interaction, cross-fertilization and synergy between researchers working on each of the above technologies. To that end, we have selected six articles, whose content we next sketch by classifying them according to the three dimensions above: three articles refer to SDC for respondent privacy, two to PPDM for owner privacy and one to PIR for user privacy.

2. Statistical disclosure control

The three papers related to SDC consider different types of data: the first of them is about the classical problem of protecting static tabular data, while the other two refer to on-line queryable databases.

The article "Adjusting the τ-Argus modular approach to deal with linked tables", by De Wolf and Giessing, is about statistical disclosure control for complex tabular data, namely linked tables. In official statistics, it is common for statistical agencies to publish multiple tabulations based on the same dataset. Those tabulations are *linked* through certain linear constraints; hence, any SDC technique aimed at respondent protection must be applied in a coordinated fashion between tables. Such a coordination can be viewed as protecting a single table of higher dimensionality, which is a computational challenge. This paper considers the secondary cell suppression technique implemented in the τ-Argus package and explores a modular approach providing respondent privacy with reasonable information loss and computational cost.

In "Regression output from a remote analysis server", O'Keefe and Good analyze the use of remote analysis servers to balance the competing objectives of allowing statistical analysis of confidential data while maintaining appropriate standards of privacy and confidentiality. Several national statistical agencies operate such servers, which do not provide data to users, but rather allow statistical analysis to be carried out remotely. A user submits a statistical query, an analysis is carried out on the original data in a secure environment, then the user receives the results of the analysis. Unless some confidentialization is put in place, the results returned to the user may leak information on the original data on which they have been computed. This article reviews results on remote analysis servers and provides a methodology for confidentializing the output from a single regression query to a remote server.

In "A Bayesian model for disclosure control in statistical databases", Canfora and Cavallo propose a novel approach for online max and min query auditing. Given a set of past max and min queries and their already disclosed answers, the query auditing system provides the answer to the current query if and only if doing so entails no privacy breach. A Bayesian network is used to assist such a decision process. The types of queries considered are substantially simpler than in the case of remote analysis servers, but the added complexity comes from taking into account the log of past queries and answers in the privacy analysis.

3. Privacy-preserving data mining

As hinted above, privacy-preserving data mining is about several data owners computing queries across their databases *without* sharing their data, but in such a way that the query issuer obtains the result of her query over the pooled set of databases. PPDM protocols are normally designed for a specific query type. The first of the two PPDM articles in this issue addresses join queries, while the second deals with frequency and classification queries.

Kantarcioglu, Inan, Jiang and Malin in "Formal anonymity models for efficient privacy-preserving joins" present a secure multiparty computation protocol that enables record joins across databanks controlled by different owners via individuals' encrypted identifiers. The authors motivate the applicability of their protocol in biomedical data sharing, a domain in which owner privacy is especially important due to the cost of data collection, the sensitivity of the collected data and the competitiveness of biomedical research.

In "Accurate and large-scale privacy-preserving data mining using the election paradigm", Magkos, Maragoudakis, Chrissikopoulos and Gritzalis discuss the design of large-scale PPDM systems in a fully distributed setting, where each client possesses its own records of private data. The article shows how to use cryptographic e-voting protocols in such a PPDM scenario.

4. Private information retrieval

The last paper in the issue, "User-private information retrieval based on a peer-to-peer community", by Domingo-Ferrer, Bras-Amorós, Wu and Manjón, is the only one addressing user privacy.

Private information retrieval for user privacy is a game between a user and a database or a search engine. The user wants to learn an item from a database without the latter learning which item is being retrieved. The authors argue that most cur-

Please cite this article in press as: J. Domingo-Ferrer, Y. Saygın, Recent progress in database privacy, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.06.002

rent PIR protocols are not practical because: (i) their computational complexity is linear in the size of the database; (ii) they assume active cooperation by the database in the PIR protocol. If the database cannot be assumed to cooperate, this article proposes to use a peer-to-peer user community to achieve some query anonymity by having the queries of a user submitted by her peers. In this way, the database still learns which item is being retrieved (hence, strict PIR is not achieved), but it cannot obtain the real query histories of users. The resulting PIR relaxation is called UPIR and a peer-to-peer UPIR system based on an underlying combinatorial structure is described.

5. Final remarks

The same ambition of cross-fertilization between SDC, PPDM and PIR which motivates this special issue is also at the root of the *Privacy in Statistical Databases-PSD* conference series sponsored by the UNESCO Chair in Data Privacy and, specifically, of the series' latest edition PSD 2008, from which the articles in this special issue were selected and extended.

For further details on current research in database privacy at the crossroads of statistics, computer science and cryptography, we recommend the proceedings of the PSD conference series; a reference to the proceedings of PSD 2008 [4] is given below.

Acknowledgement

We thank the authors for their work and also all reviewers whose insight improved the quality and the content of the papers in the issue, with a special mention to Anco Hundepool and Ercan Nergiz. We are also indebted to the late Prof. Reind van de Riet, who on behalf of DKE enthusiastically supported the idea of this special issue, but unfortunately did not live long enough to see it published.

References

[1] R. Agrawal, R. Srikant, Privacy preserving data mining, in: Proceedings of the ACM SIGMOD, ACM, 2000, pp. 439–450.
[2] B. Chor, O. Goldreich, E. Kushilevitz, M. Sudan, Private information retrieval, in: IEEE Symposium on Foundations of Computer Science (FOCS), 1995, pp. 41–50.
[3] J. Domingo-Ferrer, A three-dimensional conceptual framework for database privacy, in: Secure Data Management – 4th VLDB Workshop SDM'2007, Lecture Notes in Computer Science, vol. 4721, Springer-Verlag, Berlin, 2007, pp. 193–202.
[4] J. Domingo-Ferrer, Y. Saygın (Eds.), Privacy in Statistical Databases-PSD 2008, Lecture Notes in Computer Science, vol. 5262, Springer-Verlag, Berlin, 2008.
[5] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte-Nordholt, G. Seri, P.-P. DeWolf, Handbook on Statistical Disclosure Control (version 1.0), Eurostat (CENEX SDC Project Deliverable), 2006.
[6] J. Lane, P. Heus, T. Mulcahy, Data access in a cyber world: making use of cyberinfrastructure, Transactions on Data Privacy 1 (1) (2008) 2–16.
[7] Y. Lindell, B. Pinkas, Privacy preserving data mining, in: Advances in Cryptology – CRYPTO'00, Lecture Notes in Computer Science, vol. 1880, Springer-Verlag, Berlin, 2000, pp. 36–53.

Josep Domingo-Ferrer is a full professor of Computer Science and an ICREA-Acadèmia Researcher at Universitat Rovira i Virgili, Tarragona, Catalonia, where he holds the UNESCO Chair in Data Privacy. He received with honors his M.Sc. and Ph.D. degrees in Computer Science from the Universitat Autònoma de Barcelona in 1988 and 1991, respectively (Outstanding Graduation Award). He also holds a M.Sc. in Mathematics. His fields of activity are data privacy, data security and cryptographic protocols. He has received three research awards and four entrepreneurship awards, among which the ICREA Acadèmia Research Prize from the Government of Catalonia. He has authored 3 patents and over 220 publications, one of which became an ISI highly-cited paper in early 2005. He has been the co-ordinator of EU FP5 project CO-ORTHOGONAL and of several Spanish funded and US funded research projects. He currently co-ordinates the CONSOLIDER "ARES" team on security and privacy, one of Spain's 34 strongest research teams. He has chaired or co-chaired 9 international conferences and has served in the program committee of over 70 conferences on privacy and security. He is a co-Editor-in-Chief of *Transactions on Data Privacy* and an Associate Editor of three international journals. In 2004, he was a Visiting Fellow at Princeton University.

Yücel Saygın is a faculty member at Sabancı University, Faculty of Engineering and Natural Sciences in Istanbul, Turkey. He received his B.Sc., M.Sc. and Ph.D. degrees from the Department of Computer Engineering at Bilkent University, Ankara, Turkey, in 1994, 1996 and 2001, respectively. His main research interests include data mining and application of data mining technology to database management systems. He has done extensive research on data mining, and privacy-preserving data mining in specific. He has published in international journals like *IEEE Transactions on Fuzzy Systems, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Engineering Management*, and in proceedings of international conferences. He has co-chaired various conferences and workshops in the area of data mining and privacy-preserving data management. He also serves as a working group co-chair for the KdUbiq (Knowledge Discovery in Ubiquitous Environments) Coordination Action project funded by the European Commission.

Please cite this article in press as: J. Domingo-Ferrer, Y. Saygın, Recent progress in database privacy, Data Knowl. Eng. (2009), doi:10.1016/j.datak.2009.06.002