Challenges for privacy preservation in data integration

PETER CHRISTEN and DINUSHA VATSALAN, The Australian National University
VASSILIOS S. VERYKIOS, Hellenic Open University

Techniques to integrate data from diverse sources have attracted significant interest in recent years. Much of today's data collected by businesses and governments are about people, and integrating such data across organizations can raise privacy concerns. Various techniques that preserve privacy during data integration have been developed, but several challenges persist that need to be solved before such techniques become useful in practical applications. We elaborate on these challenges and discuss research directions.

Categories and Subject Descriptors: H2.8 [**Database Management**]: Database Applications-*Data Mining*

General Terms: Design, Algorithms, Security

Additional Key Words and Phrases: Privacy techniques, privacy-preserving record linkage, data matching

ACM Reference Format:
Peter Christen, Dinusha Vatsalan, and Vassilios S. Verykios, 2014. Challenges for privacy preservation in data integration. *ACM J. Data Inform. Quality* 99, 99, Article 99 (Month 2014), 2 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Integrating data across organizations can lead to tremendous benefits, for example to improve data quality or allow the discovery of new and valuable knowledge that is not available from individual databases [Doan et al. 2012]. Data integration includes three major aspects: *schema matching* [Bellahsene et al. 2011], where the aim is to identify which attributes and tables in several databases contain the same type of information; *data matching* [Christen 2012], also known as record linkage or entity resolution, where the aim is to identify which records in one or more databases refer to the same entity; and *data fusion* [Bleiholder and Naumann 2008], the task of merging records that refer to the same entity into consistent and coherent forms.

When data about individuals, or otherwise sensitive data, are to be integrated across organizations, privacy and confidentiality have to be considered and these data need to be protected from unauthorized disclosure. Domains where privacy preservation during data integration is important include health services, business collaborations, national censuses, the social and health sciences, crime and fraud detection, and national security. Increasingly, applications in these domains require data to be integrated from sources both internal and external to an organization. Collecting sensitive data in one location for integration and analysis makes them vulnerable to both external and insider attacks, as recent national security data leakages have shown. It would be much better if sensitive data could be kept at their sources, while still allowing them to be integrated and analyzed without revealing any private or confidential information.

Supported by the Australian Research Council under Discovery Project DP130101801. Author's addresses: P. Christen and D. Vatsalan, The Australian National University, Canberra, Australia; V.S. Verykios, Hellenic Open University, Patras, Greece; email (corresponding author): peter.christen@anu.edu.au

DOI: http://dx.doi.org/10.1145/0000000.0000000

Only limited work has investigated privacy in the context of schema matching [Scannapieco et al. 2007], where the aim is to develop methods that do not reveal any sensitive information about the source schemas and data. Techniques that consider privacy within data matching are known as privacy-preserving record linkage (PPRL). The aim of PPRL is to match records across databases in such a way that besides the matched records (those classified to refer to the same entities) no information about the source data can be learned by any party involved in the linkage, or any external party. As surveyed in Vatsalan et al. [2013], nearly thirty approaches for PPRL have been developed. The principle requirements in PPRL are to allow for approximate matching of values to overcome data quality issues; to have techniques that are invulnerable to any kind of attack; and that are also scalable to matching large databases. In the context of information systems, no work has investigated privacy-preserving data fusion.

2. CHALLENGES AND RESEARCH DIRECTIONS

In order to achieve fully practical privacy-preserving data integration, the following major challenges need to be further addressed and eventually solutions be found.

(1) Current approaches for PPRL have limited scalability towards very large databases, and can only match static data in batch mode. Novel techniques that facilitate efficient indexing (identifying groups of records to be compared in detail), real-time matching, and that can handle dynamic data, need to be developed.

(2) In a PPRL framework the use of training data for supervised classification of compared records into matches and non-matches is challenging, as the actual attribute values need to be known for training. Advanced unsupervised approaches, such as collective and graph-based classification techniques that have been developed for data matching, need to be investigated within a privacy-preserving framework.

(3) Assessing the quality and completeness of integrated data without revealing any sensitive information is difficult. However, not knowing how accurate and complete integrated data are renders any privacy-preserving data integration approach impractical. Recent work on interactive assessment of linked records [Kum et al. 2013] is promising but not scalable to large databases. Large-scale assessment can potentially be achieved by using synthetic data that are generated based on real data.

(4) Many real-world applications require data from more than two sources to be integrated. Most current privacy-preserving data integration techniques however have only considered data from two sources. Besides computational challenges, the issue of collusion between (sub-sets of) parties needs to be carefully considered.

(5) Developing techniques for privacy-preserving data fusion could lead to new applications that allow data quality improvements for sensitive and confidential data.

In summary, integrating data can lead to significant benefits, however privacy concerns often need to be considered. Various challenges still need to be addressed before scalable, accurate, and privacy-preserving data integration becomes practical.

REFERENCES

Zohra Bellahsene, Angela Bonifati, and Erhard Rahm. 2011. *Schema matching and mapping*. Springer.

Jens Bleiholder and Felix Naumann. 2008. Data fusion. *Comput. Surveys* 41, 1 (2008), 1–41.

Peter Christen. 2012. *Data matching - Concepts and techniques*. Springer.

AnHai Doan, Alon Halevy, and Zachary Ives. 2012. *Principles of data integration*. Morgan Kaufmann.

Hye-Chung Kum, Ashok Krishnamurthy, Ashwin Machanavajjhala, and Michael K. Reiter Stanley Ahalt. 2013. Privacy preserving interactive record linkage (PPIRL). *J. Am. Med. Inform. Assoc.* (2013).

Monica Scannapieco, Ilya Figotin, Elisa Bertino, and Ahmed K. Elmagarmid. 2007. Privacy preserving schema and data matching. In *ACM SIGMOD*. Beijing, 653–664.

Dinusha Vatsalan, Peter Christen, and Vassilios S. Verykios. 2013. A taxonomy of privacy-preserving record linkage techniques. *Information Systems* 38, 6 (2013), 946–969.