Effective Preservation of Privacy During Record Linkage

[Extended Abstract]

Yang Lu
University of Melbourne
luy4@student.unimelb.edu.au

Richard O. Sinnott
University of Melbourne
rsinnott@unimelb.edu.au

Karin Verspoor
University of Melbourne
karin.verspoor@unimelb.edu.au

Udaya Parampalli
University of Melbourne
udaya@unimelb.edu.au

ABSTRACT

Record linkage is a technique for integrating data from potentially multiple sources where direct access to data is not allowed due to security and privacy considerations. To avoid privacy leakage from and over-modifications of raw data, we propose *semantic-based linkage k-anonymity* (SLA) as an effective solution to de-identifying clinical record linkage.

CCS Concepts

• Information systems → Database management system, Data mining; • Security and privacy → Human and societal aspects of security and privacy.

Keywords

Record linkage; k-anonymity; semantic reasoning

1. INTRODUCTION

Record linkage has been recognised as a key approach to support in-depth research on areas including public health and individual well-being. For instance, the Centre for Health Record Linkage (CHeReL, http://www.cherel.org.au/) relies on probabilistic matching of demographic data to create linked health records across the New South Wales and Australian Capital Territory. Due to the sensitivities of individual records, linkage needs to check privacy risks before releasing data to applicants. To avoid latent risks caused by skewed attributes, we design a semantic model to preserve individual privacy while reducing information loss by data modifications (e.g. generalisation).

2. METHOD

Sweeney (2002) proposed that *k-anonymity* processing quasi-identifiers (QIs) of data may lead to 'over generalisation' when dealing with linkage data sets [1]. Based on the fact that most linkage cases do not include all local patients and thus not all modifying data for privacy-preserving purposes needs to be used, we propose the *linkage k-anonymity* (LA) by which only obfuscated individuals in a released linkage set are required to be indistinguishable from at least *k-1* other individuals in the local dataset. To the inference disclosure issue, however, the general solutions involve ruling out risky associations from previous linked data releases. Originally applied to discover associations from transaction records [2], the association rule mining approach is adopted to deal with anonymised linkage. As a consequence, the *semantic-based linkage k-anonymity* (SLA) workflow contains functional components used to satisfy both requirements (shown in Figure 1). Specially, associations identified from the 'previous release' become the input of semantic reasoning for 'current linkage'.

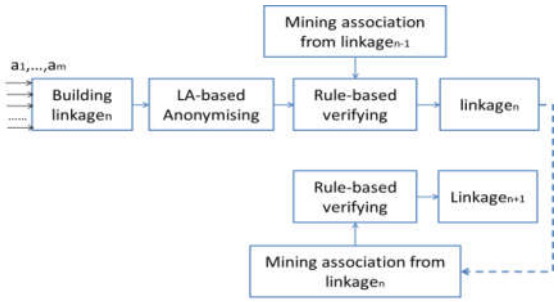

Figure 1. Workflow in SLA

3. EXPERIMENT

The scenario underpinning this work uses data collected from 1000 patients involved in the Australasian Diabetes Data Network (http://www.addn.org.au/) and 1850 individuals from the VicHealth population survey (https://www.vichealth.vic.gov.au/) undertaken in 2015. The experiment considers the case where 500 individuals are in both data sets. After using *LA* and *SLA* on linkage sets, the performances are evaluated by comparing the average disclosure risk and sum of squared error (SSE) respectively [3]. The result shows *LA* is able to preserve more details while it runs a higher risk of re-identification than using *SLA*. Although the verification of *SLA* results in more data generalisation, the major improvement in data quality is due to the linkage *QI* attribute filter.

4. CONCLUSION

We designed an approach for *semantic-based linkage k-anonymity* (*SLA*) based on linkage properties. In the future, we will explore semantic approaches for privacy preserving record linkage (PPRL) where higher accuracy linkage is expected.

5. REFERENCES

[1] L. Sweeney. 2002. *k-anonymity: A model for protecting privacy.* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-570. DOI= http://doi.acm.org/10.1142/S0218488502001648

[2] Agrawal, R., Imieliński, T., & Swami, A. 1993. *Mining association rules between sets of items in large databases.* ACM SIGMOD Record. 22(2), 207-216. ACM. DOI= http://doi.acm.org/10.1145/170035.170072

[3] J. Domingo-Ferrer, A. Martínez-Ballesté, J. M. Mateo-Sanz & F. Sebé. 2006. *Efficient multivariate data-oriented micro-aggregation*. The VLDB Journal—The International Journal on Very Large Data Bases, 15(4), 355-369. DOI= http://doi.acm.org/10.1007/s00778-006-0007-0