

Improved record linkage for encrypted identifying data

Chaoyi Pang, David Hansen

E-Health Research Centre, ICT Centre, CSIRO, Brisbane, Australia

ABSTRACT

*The health data integration project at the E-Health Research Centre is researching ways of improving the integration of health and health related data while maintaining the privacy and security of the data. One such method is to improve the mechanisms of matching patients across databases when the identifying information must not be revealed, even during the linkage step. **Background:** With health related data spread between many administrative and clinical databases the ability to bring the data together dynamically is important. This could be to support clinical based decision making, administrative reporting or for clinical research based access to data. **Objectives:** There are already mechanisms published for blind folded record linkage. A mechanism for further strengthening the security and privacy of these algorithms is to encrypt the identifying data, such as name, data of birth, before performing the linkage step. However, due to the nature of encryption algorithms, encrypted data can only be matched exactly, limiting the ability to allow for errors in the data. This work presents a mechanism to allow matching of encrypted data when there may be errors in the data. **Methods:** A public reference table which is common to both data custodians is used. Each value in the original data is compared to data in the public reference table using an edit distance function. Names from the reference table which are within a given distance of the original data are sent to the linker. The data from the two data custodians are then compared to decide the likelihood of two records being a match. **Results:** The method described in this paper performs better than other methods which support matching of encrypted data, such as exact matching or matching using soundex. **Discussion and Conclusion:** The method described in this paper can be used to improve the level of record matching in tools where access to identifying data is prohibited. This method is currently being added to the HDI software tool as another mechanism of matching records between databases.*

Keywords:

Medical Record Linkage; Confidentiality; Privacy;

INTRODUCTION

A common requirement of applications in public health and biomedical research is the ability to link records in disparate databases which refer to the same person. This is particularly the case in Australia's health system where data about a patient is spread between many custodians; and there is no common identifier used throughout the system.

A number of such record linkage techniques have been proposed in the last two decades [1, 2, 3]. The E-Health Research Centre is researching new methods for record linkage for inclusion in a software tool, HDI [4, 5]. HDI is a tool for the integration of health and health related data while maintaining the privacy and security constraints incumbent on these data. The HDI software offers a multi-party system for integrating data sources. Data resides within the data custodian's computer and a data service is used to access the data. A linking service acts independently to perform the matching step and allow for the clinical data related to an individual to be linked without revealing their identity.

In line with best practice, the linking service does not have access to the non-identifying (e.g. clinical) data when matching patients from different data sources [6, 7]. Despite this, the linking service could infer knowledge about an individual from the fact that they have data in a particular data source. To mitigate this, HDI encrypts the identifying data sent to the linking service for matching, using a one-way keyed hash function [8]. The configurable linking algorithm used by HDI allows data custodians to specify the identifying fields (name, date of birth, age, etc) which are available for linking and how they relate to a standard data dictionary. The linking service then infers probabilistically whether records refer to the same patient.

While the encryption of identifying data offers extra data security and patient privacy, it limits the ability of the linking service to correct for spelling or typographical errors. Phonetic encoding functions such as soundex or double metaphone can be used to transform original strings in data sources; however, these transformations are not robust with respect to errors in the initial character or to truncation differences. Furthermore, they can also lead to a large increase in the number of false matches.

In this paper, we report a practical method of privacy-preserving approximate string matching via a public reference table. As shown in the results section, this method of matching offers a large improvement in matching over direct matching of encrypted strings and phonetic encodings.

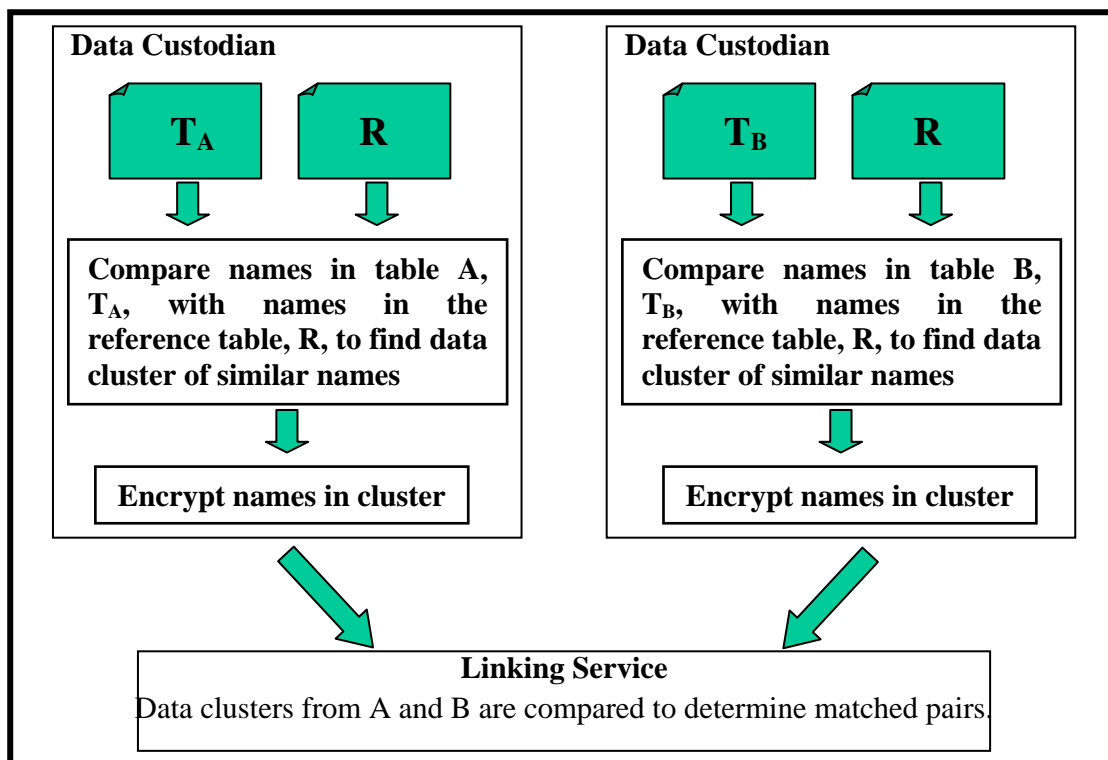
METHOD

The algorithm described in this paper is to allow comparison of strings when the strings must be encrypted. Direct comparison of the strings allows for an exact match only, meaning there can be no allowance made for errors in the data. This algorithm compares the two strings indirectly, by instead comparing cluster of values selected from a publicly available table.

The approximate string matching algorithm can be described by 4 steps:

1. a cluster of values for each string is identified from the public reference tables using an edit distance function.;
2. each string in this cluster is encrypted
3. the data custodian then sends the encrypted string values, along with their distances, to the linking service;
4. the linking service calculates a matching score from the information sent from the two data sources.

When performing matching in our setting, the encrypted reference data from the reference table is sent to the linking service together with associated distance values. Not sending the actual encrypted data improves data privacy as the actual data does not leave the data custodian, even in an encrypted form, and is thus less available to other parties.



Finding similar strings in the reference table

An edit distance function is used to identify a cluster of strings, or neighbourhood, from the reference table for each string. Each string is compared to every string in the reference table using an edit distance function. If the edit distance is less than a threshold, δ , then the string in the reference table is added to the cluster.

The reference table is a collection of unique strings reflecting the domain of the strings to be matched. In the case of names, a table of unique names from the local phonebook or electoral roll can be used.

Encrypting and transferring comparison results

Because of the data confidentiality requirement, all sensitive fields, such as names and addresses, are encrypted before they are sent outside the original data custodian. In order to make comparison at the linking service, a secret encryption-key must be generated and shared among custodians (and must be unavailable to the linking service) [7].

After the encryption is completed, each data custodian sends the values in the cluster, along with their distance from the original value to the linking service.

Calculating similarity scores at the linking service

We would expect clusters will overlap if a pair of names is similar. The intersection of the two clusters can be found by comparing the encrypted values (equality) of names in the two clusters. Using the triangle inequality property, we can calculate the upper bound of the distance between the name pair via each of the names in the intersection region. If the minimum distance is less than a certain threshold, δ_{sim} , the string pair is said to match.

EXPERIMENTAL RESULTS

To evaluate the performance of this algorithm, we compared this algorithm against other methods which allowed the data to be encrypted. The experiments performed consisted of finding matches between two data sets with a known number of matches. We used datasets which are part of the

FEBRL package [9] as the test data sets. FEBRL is a tool used for approximate matching of people between datasets using a variety of mechanisms. In the experiments we used different reference tables and edit distance thresholds.

Test datasets

For our experiment we used the test data sets **4a** and **4b** from the FEBRL package. The datasets contain a range of synthetic identifying information (names, address, etc) for the same 5000 people, with errors introduced into both datasets. We have removed just over 500 duplicate records and records with a missing surname or given name, to give a final test data set of the same 4468 records.

We conclude that two records are a match if both the surname and given name meet the criteria for a match in our experiment and the match is correct if the record IDs are the same. Where we have used the public reference table, 2 names are a match if the smallest edit distance between values in their clusters is less than the threshold distance, δ_{sim} .

Reference tables

To examine the importance of the reference table for achieving the best results, we use four reference tables in the experiments. The first reference table used is the combined unique given names and surnames from the first data set, **4b**, giving a reference table size of 2544 unique names. The second reference table is every second name from the first reference table, starting with the first name, while the third reference table is every second name starting with the second name.

The fourth reference table contains unrelated data – the unique surnames and given names from the Internet Movie Database [10]

Performance metrics

To evaluate the performances of our method under different parameter settings, the metrics of *precision* and *recall* were used, as these metrics are commonly used in information retrieval [11]. Precision refers to the accuracy of a name matching method, and in this case is defined as the percentage of correctly matched name pairs among all matched name pairs. Recall is defined as the percentage of correctly matched name pairs among all true matched name pairs.

Results

Table 1 details the results of the experiments. As well as our algorithm, we have used two other mechanisms for finding matches - direct comparison of the names and comparison of the soundex of the names.

Method	δ_{sim}	Correct Matches	Incorrect Matches	Precision	Recall
Direct comparison	-	1977	0	1.0	0.39
Soundex	-	3246	774	0.76	0.55
Reference Table 1	2	3332	56	0.98	0.73
Reference Table 2	2	907	28	0.97	0.20
Reference Table 3	2	911	23	0.97	0.20
Reference Table 4	2	1404	38	0.97	0.30
Reference Table 1	3	4767	765	0.84	0.89
Reference Table 2	3	1911	384	0.8	0.34
Reference Table 3	3	1828	335	0.82	0.33

The best results obtained using our test data when using the new algorithm with the 1st reference table and a δ_{sim} of 2. Using this algorithm shows both better precision and recall than phonetic

encoding and the other reference tables. Reference table 1 is a super set of likely names in the data to be matched, showing the importance of the reference table to this algorithm. Ideally, the reference table should be a super set of names likely to be in the data to be matched.

CONCLUSIONS

This new algorithm provides a new way of matching on even the most sensitive of data while still allowing for errors in the data, providing a new way of matching on identifying string values. The configurable HDI linking algorithm contains other mechanisms for matching on numerical data and when used in conjunction with this new method will provide a more complete linking algorithm for sensitive data.

REFERENCES

- [1] P. Christen, T. Churches and M. Hengland. Febrl – A parallel Open Source Data Linkage System. In *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04)*, pages 638-647, Sydney, Australia, May 2004.
- [2] W. Winkler. The state of record linkage and current research. In *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pages 73-80, 1999.
- [3] W. Cohen. Data integration using similarity joins and a word-based information representation language. *ACM Transactions in Information Systems*, 18(3):288-321, 2000.
- [4] K.L. Taylor, C.M. O'Keefe, J. Colton, R. Baxter, R. Sparks, U. Srinivasan, M. A. Cameron L. Lefort, "A Service Oriented Architecture for a Health Research Data Network", Proc. SSDBM '04
- [5] D. Hansen, C. Daly, K. Harrop, M. O'Dwyer, C. Pang, and J. Ryan-Brown. HDI: Research Software To Commercial Product, ASWEC 2005 Industry Experience Papers.
- [6] L.G. Christine M OKeefe, Ming Yung and R. Baxter. "Privacy-preserving linkage and data extraction protocol. In *Workshop on Privacy in Electronic Society in Conjunction with the 11th ACM CCS meeting*, Washington DC, 2004
- [7] T. Churches and P. Christen. Some methods for blindfolded record linkage. *BMC Medical Informatics and Decision Making*, 4(1):0, 2004.
- [8] L. Dusserre, C Quantin and H. Bouzelat. A one way public key cryptosystem for the linkage of nominal files in epidemiological studies. *International Journal of Medical Informatics*, 8:644-647, 1995.
- [9] <http://sourceforge.net/projects/febrl> last accessed 10th March 2006.
- [10] The Internet Movie Database <http://www.imdb.com/> last accessed 13th March 2006.
- [11] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley Professional, 1999.

Lead Author Contact: Dr Chaoyi Pang, email: Chaoyi.pang@csiro.au phone: (07) 3024 1611