LinkIT: Privacy Preserving Record Linkage and Integration via Transformations

Luca Bonomi
Math&CS Department
Emory University, Atlanta, GA
lbonomi@emory.edu

Li Xiong
Math&CS Department
Emory University, Atlanta, GA
lxiong@mathcs.emory.edu

James J. Lu
Math&CS Department
Emory University, Atlanta, GA
jlu@mathcs.emory.edu

ABSTRACT

We propose to demonstrate an open-source tool, LinkIT, for privacy preserving record Linkage and Integration via data Transformations. LinkIT implements novel algorithms that support data transformations for linking sensitive attributes, and is designed to work with our previously developed tool, FRIL (Fine-grained Record Integration and Linkage), to provide a complete record linkage solution. LinkIT can be also used as a stand-alone secure transformation tool to link string records. The system uses a novel embedding technique based on frequent variable length grams mined from original records with differential privacy, and utilizes a personalized threshold for performing linkage in the embedded space. Compared to the state-of-the-art secure transformation method [16], LinkIT guarantees stronger privacy with better scalability while achieving comparable utility results.

Categories and Subject Descriptors

H.2.7 [**Database Management**]: Database Administration—*Security, integrity, and protection*

Keywords

Privacy, Security, Record Linkage

1. INTRODUCTION

The continuing need for data integration and analysis from multiple data sources has drawn increasing attention to the problem of record linkage [8, 18]. The record linkage process consists in identifying records that refer to the same real world entity across different sources. With the increasing amount of personal information being collected and linked, it is crucial that the *sensitive* information in the data is not disclosed during the linkage process. For example, in a decentralized healthcare system, where the personal medical records are distributed among several hospitals, it is critical to integrate the information of a patient without disclosing his/her sensitive attributes. This poses the challenge of providing accurate linkage results while simultaneously guaranteeing privacy and security in the linkage process.

SIGMOD'13, June 22–27, 2013, New York, New York, USA.

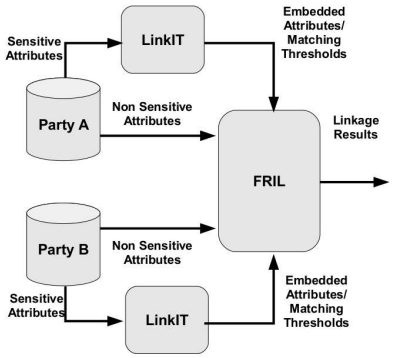


Figure 1: Integrated solution overview

Several techniques have been proposed in the literature to perform record linkage in a privacy preserving and secure way. They are generally based on Secure Multiparty Computation (SMC) [15, 20], secure transformation [1, 5, 16, 17], and hybrid methods [9, 11, 13]. SMC techniques use cryptographic mechanisms and allow two parties to perform record linkage as a secure function such that no party knows anything except its own input and the results. However, they are computationally prohibitive in practice. Secure transformation methods use data transformation techniques such as one-way hashing or embedding to map the original data into new data values that cannot be reversed and then perform record linkage on the transformed data, typically by a third party. While these methods are more efficient, the challenge is to have a secure transformation while preserving the accuracy of linkage on the transformed space as high levels of protection typically imply a great loss of accuracy in the final results. Finally, hybrid techniques attempt to combine anonymization or transformation techniques with SMC protocols. They typically use a privacy preserving *blocking* step to restrict the comparisons to smaller groups of records which are then matched by SMC protocols. While they provide a trade-off between efficiency and accuracy, the SMC step is still required and is typically not implemented or evaluated due to the high computation cost.

We demonstrate LinkIT, a system for privacy preserving record Linkage and Integration via data Transformations. LinkIT implements novel algorithms that support data transformations for linking sensitive attributes. It is designed to work cooperatively with our previously developed tool, FRIL (Fine-grained Record Integration and Linkage), to provide a complete record linkage solution. An overview of an integrated solution in a two-party scenario is illustrated in Figure 1. LinkIT can be employed by each data party

holder to transform the original sensitive attributes to embedded attributes and generate matching thresholds to be used for matching in the embedded space. A third party (not necessarily trusted), in this case FRIL, performs the linkage on the original non-sensitive attributes and the embedded sensitive attributes using the matching threshold in the embedded space.

The LinkIT module implements and extends the embedding approach we have recently proposed [2, 3]. It uses a novel embedding scheme based on frequent variable length grams. LinkIT can run as a stand-alone software for linking string attributes or serve as a preprocessing module to FRIL to provide a complete record linkage toolkit for a variety of data attributes. FRIL [12] is an open source software that we have developed for a birth defects surveillance program at the CDC (Center for Diseases Control and Prevention). It has since been adopted by many institutions and the code and tutorials are available at http://fril.sourceforge.net/.

The primary contribution of LinkIT is a secure transformation technique with several novel features for linking sensitive string attributes. It uses an embedding strategy based on frequent variable length grams that are mined from the original data. This choice is motivated by the fact that the strings being matched in record linkage scenarios typically have similar properties (e.g. same alphabet, similar length, etc.), so this gives an important advantage compared to a randomly generated base for embedding. The frequent grams are mined from the original records under the formal and provable guarantees of *differential privacy* [6], hence providing a strong privacy protection. Furthermore, LinkIT uses a novel concept of *personalized threshold* [3] that significantly improves traditional strategies based on a global threshold for matching data in the embedded space. Our approach dynamically computes a personalized threshold for each record, which represents the maximum distance between potential matching records in the embedded space. This differs from the global threshold scheme that requires an a priori global threshold. In addition to improving accuracy of the linkage results, the mechanism enhances the portability of our system and minimizes the manual parameter tuning required from the users.

2. SYSTEM DESCRIPTION

We will give an overview of the integrated system first and then describe the LinkIT and FRIL in detail. The original data are classified into sensitive and non-sensitive sets according to the specification provided by the user. The sensitive attributes are fed into LinkIT that provides a private and secure transformation mechanism so that they can be linked using the linkage module. This transformed data together with the non-sensitive attributes are handled by FRIL. The overall system receives the initial files containing the records, and a set of parameters to define the similarity measure and distance functions for matching the attributes. As an output, the system produces three sets: M (matched records), U (unmatched records), P (possible matches) and various summary statistics. The sets U and P may be fed back into the linkage module. In the rest of the section, we describe the details of the components and algorithms used in LinkIT and FRIL.

2.1 LinkIT Components

LinkIT implements several novel features. First, it uses an embedding strategy based on frequent variable length grams which achieves a considerable gain in result accuracy with respect to a random base. Second, the frequent grams are mined using a privacy preserving miner, so that no sensitive information of the individual records is disclosed. Finally, it provides a personalized threshold scheme that allows the third party to effectively perform

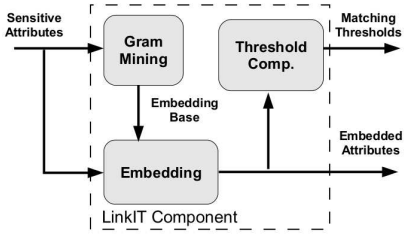
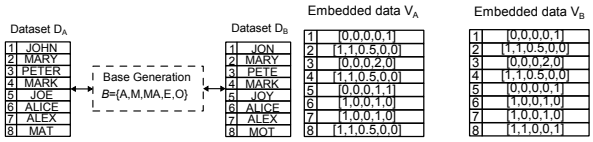


Figure 2: LinkIT overview



(a) Mining and Base Generation (b) Embedded Data

Figure 3: Example of Mining, and Embedding of the data.

approximate matching in the embedded space. We consider two string records with an edit distance [14] smaller than a user specified threshold ed as a match. In addition to the global threshold matching which allows the user to specify a threshold value, LinkIT dynamically computes a personalized threshold value for each record in the embedded space without requiring any extra information a priori.

An overview of the LinkIT components and its conceptual flow is illustrated in Figure 2. LinkIT receives sensitive strings such as names as input, and produces as output a set of personalized threshold values to use in the linkage module and a set of embedded vector data representing the input data. It consists of several components: mining, embedding, and threshold computation. The user can select the mining method to use in generating the base for the embedding, specify the privacy level, and the matching threshold in the original space (value of ed). In the rest of the section, we briefly describe the major components in LinkIT. For further details regarding our novel technique, we refer the readers to [2, 3].

Mining. The Mining component mines the original records for the top-k *frequent* variable length grams, where a gram is a substring of the original strings. The resulting grams are used to form an embedding base for embedding the original strings. Contrary to existing transformation techniques, LinkIT uses an embedding base mined from the original data so that the original records can be better represented by the embedding process. In order to guarantee that no sensitive information is disclosed in the mining step, we adopt the differential privacy framework [6]. This framework has been widely accepted in recent years as a strong and provable privacy guarantee. It requires that the outcome of a computation to be indistinguishable when run on datasets with presence or absence of any individual record. Figure 3(a) shows an example of two parties holding a set of string records, D_A and D_B, respectively. {A, M, MA, E, O} are the combined frequent grams mined from the input data that satisfy differential privacy for individual string records in the input data. Note that the two parties can exchange the resulting frequent grams since they do not disclose individual records. The grams will then be used as the embedding base for the subsequent embedding step.

LinkIT includes a privacy preserving mining algorithm which uses the Laplace mechanism [7] to guarantee differential privacy

for the mining step. The required input from users includes the maximum length of the grams, the number of grams to be mined or the size of the base k, and the privacy level ϵ (privacy budget), which will impact the quality of the base due to the level of noise perturbation. The mining algorithm is an extension of the prefix tree approach used for mining trajectory data in [4]. We proposed a variety of techniques to distribute the privacy budget in the tree that enhance the performance of the original method as shown in [2].

Embedding. Once an embedding base is constructed using the top-k frequent grams, the Embedding component transforms each string into a vector in the space $\mathbb{R}^k$. For each embedded vector, the i-th component in the vector represents the number of occurrences of the i-th gram of the base in the original string, scaled by the length of the gram itself. An example of embedded data is illustrated in Figure 3(b). The challenge of an embedding approach is how to select a proper matching threshold to ensure that the matching records in the original space can be matched in the embedded space. In our embedded space, the distance between two vectors is computed using the Euclidean distance. We will explain the computation of the matching threshold in next subsection.

As an alternative embedding method, the Lipschitz embedding approach [16] has been recently proposed. The main limitation of that approach is that the embedding base is either formed by random strings that impacts the linkage accuracy, or optimized based on the original records at one party. The latter incurs disclosure risks of the sensitive information. In contrast, our projection based embedding is based on the frequent grams from original records with differential privacy and hence provides a formal privacy guarantee while offering an accurate representation of the original strings.

Threshold computation. In the original space, we are interested in matching strings within ed edit operations. In the new space, this task becomes finding all vectors whose Euclidean distance is within a certain threshold. This threshold value plays a central role on the overall performance and is computed by the Threshold Computation component. In addition to the option of computing a global threshold [2], LinkIT also provides a novel personalized threshold scheme [3]. The choice of using a personalized threshold is motivated by the following reasons. First, each string shares a different number of grams with the base, and for those strings that have a large number of shared grams, a personalized threshold can better represent the original distance. Second, by computing a threshold for each string, we overcome the problem of estimating a threshold suitable for all strings which could be hard. To compute this personalized threshold, we use a dynamic programming algorithm inspired by the work in [19]. For each string s, and the set N of neighboring strings based on the embedding base within ed edit distance; the personalized threshold is the maximum distance between the embedded vector of s and the set of embedded vectors of N.

Other sensitive attribute types. Our current prototype also has some preliminary solutions to handle sensitive non-string data. One possibility consists in converting these attribute values into strings first, and applying the same strategy to link them. In this case, the similarity measure between these converted attributes can be computed by a weighted edit distance. A second solution provided by LinkIT is based on *suppression* and *encryption*. The suppression process consists in removing a certain attribute from the records. This strategy has to be carefully applied since it may lead to poor performance in the linkage due to the lack of critical linking fields. While encryption technique provides high security and supports exact match on the encrypted data, it remains a challenge to support effective approximate matching.

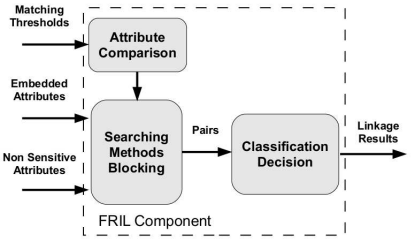


Figure 4: FRIL overview

2.2 FRIL components

FRIL adopts the probabilistic linkage approach in [10], with enhanced control and tuning options for the user as well as automatic schemes. In our integrated solution, FRIL is used to perform the record linkage on the non-sensitive data as well as the sensitive data transformed by LinkIT. The set of personalized thresholds produced by LinkIT will be used to match the embedded data. In addition, the user can specify several parameters such as attribute weights and threshold values for linking various attributes. Figure 4 shows the major components in FRIL: searching or blocking, attribute comparison, and classification/decision. The searching component specifies how the data are compared by allowing one-to-one or blocking comparison between the records. The classification component allows user-defined importance values for each attribute on the final matching result as well as provides semi-supervised classifiers for matching. For more details regarding FRIL, we refer the readers to [12] and the FRIL project site[1].

3. DEMONSTRATION

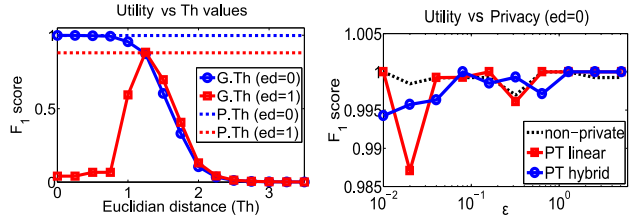
We plan to demonstrate the novelty and utility of the LinkIT system both as a stand-alone tool and as a preprocessing module of FRIL. As a stand-alone tool, we will demonstrate how sensitive string records can be efficiently and effectively matched using the secure transformation approach. As a complete system, we will demonstrate a running example of our record linkage toolkit to match more complex type of records with more user interactions.

Datasets. To demonstrate LinkIT as a stand-alone tool for linking sensitive strings, we use two real datasets, `NAMES`[2] and `CITIES`[3]. The first contains a list of the most frequent surnames from the Census 2000. The second is a list of the top 5000 most populated cities in U.S. in 2008.

To demonstrate the record linkage toolkit, we will first use some synthetic data. In addition, to test our framework on real scenario we plan to use the `Adult` dataset from the UCI Machine Learning Repository[4].

LinkIT Functionality. First, we will demonstrate the functionality of LinkIT for sensitive string data (e.g. names). Through a graphical user interface, conference audience can select the dataset, the desired level of differential privacy ϵ for the mining phase, the desired size for the embedding base, the similarity metric and threshold. The system will then display a sample list of linked strings, and the utility of the linked results in term of F_1 score which is

[1] `http://fril.sourceforge.net/`
[2] `http://www.census.gov/genealogy/www/data/`
[3] `http://www.census.gov/popest/data/`
[4] `http://archive.ics.uci.edu/ml/`



(a) **Utility vs Global Threshold** (b) **Utility vs Privacy**

Figure 5: Utility Results

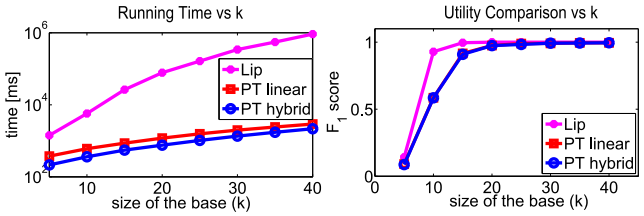


Figure 6: Performance: (Left) Running time, and (Right) Utility for Lipschitz and Frequent gram embedding (PT linear, PT hybrid).

a combination of precision and recall. These parameters play an important role in the overall utility results. Users may vary the parameters, or select different setting for the mining algorithm and embedding method, and examine the different outputs and accuracy of the linkage results.

To demonstrate the benefit of our personalized threshold component in LinkIT, we will compare the global threshold schema and personalized schema. In the demo, users can specify different values to be used as a global threshold and compare the results with those obtained using the personalized threshold. For example, Figure 5(a) shows that the utility with the personalized threshold (dashed line) is as high as the best results obtained by setting appropriate value of global threshold on the `NAMES` dataset. These show the advantages of using the personalized threshold strategy which computes an optimal value of threshold without requiring a priori knowledge.

The audience can specify different differential privacy values and Figure 5(b) illustrates the dependency between privacy level (ϵ) and the utility in linking string records using our embedding strategy on the `CITIES` dataset. Two variants of the prefix-tree (PT) based mining algorithms, PT linear and PT hybrid, are compared to a baseline non-private approach. The utility increases as the privacy level decreases (larger value of ϵ), and approaches the result obtained by a miner that uses a non-private mining algorithm. This shows the great quality of the matching results provided by LinkIT.

The Lipschitz strategy [16] can be used as a comparison in the embedding step. Figure 6 shows that our strategy produces similar utility results with a significant lower computational cost compared to the Lipschitz embedding.

Interface and Integrated Record Linkage. We will also demonstrate the integrated system for linking a variety of attributes. Users can specify the sensitive and non-sensitive attributes. The integrated system will seamlessly suppress or transform the sensitive attributes through LinkIT and match the different attributes through FRIL. FRIL provides a main interface window, where the user can configure and visualize the data sources, as well as set the parameters for the linkage step. By selecting the data sources, a user can perform several actions on each individual record (e.g. visualization, conversion and merging/split of attributes). The user may directly control the linking procedure by selecting several distance functions between attributes and obtain a real-time visualization of the results. In this way, FRIL supports fine-grained and flexible user decision and allows the user to better understand the impact of each parameter on the linkage results. For users who prefer automatic schemes, FRIL also provides automatic parameter settings and semi-supervised learning schemes. When the results are reported, the user may refine the result by repeating the linkage process on sets U and P, adjusting the decision and classification methods. When a desired output is obtained, the final results are saved.

4. ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1117763.

5. REFERENCES

[1] A. Al-Lawati, D. Lee, and P. McDaniel. Blocking-aware private record linkage. IQIS '05. ACM, 2005.

[2] L. Bonomi, L. Xiong, R. Chen, and B. C. M. Fung. Frequent grams based embedding for privacy preserving record linkage. In *CIKM*. ACM Press, 2012.

[3] L. Bonomi, L. Xiong, R. Chen, and B. C. M. Fung. Privacy Preserving Record Linkage via grams Projections. *ArXiv e-prints:1208.2773*, 2012.

[4] R. Chen, B. C. M. Fung, B. C. Desai, and N. M. Sossou. Differentially private transit data publication: A case study on the montreal transportation system. ACM Press, 2012.

[5] T. Churches and P. Christen. Some methods for blindfolded record linkage. *BMC Medical Informatics and Decision Making*, 2004.

[6] C. Dwork. Differential privacy. In *in ICALP*, 2006.

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC 2006*, 2006.

[8] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Trans. Knowl. Data Eng.*, 2007.

[9] L. O. Evangelista, E. Cortez, A. S. da Silva, and W. M. Jr. Adaptive and flexible blocking for record linkage tasks. *JIDM*, 2010.

[10] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 1969.

[11] A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino. Private record matching using differential privacy. EDBT '10, 2010.

[12] P. Jurczyk, J. J. Lu, L. Xiong, and J. D. Cragan. Fril: A tool for comparative record linkage. In *In AMIA Annual Symposium*, 2008.

[13] M. Kuzu, M. Kantarcioglu, A. Inan, E. Bertino, E. Durham, and B. Malin. Efficient privacy-aware record integration. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, pages 167–178, New York, NY, USA, 2013. ACM.

[14] V. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.

[15] Y. Lindell and B. Pinkas. Secure multiparty computation for privacy-preserving data mining. Cryptology ePrint Archive, Report 2008/197, 2008.

[16] M. Scannapieco, I. Figotin, E. Bertino, and A. K. Elmagarmid. Privacy preserving schema and data matching. SIGMOD '07. ACM, 2007.

[17] R. Schnell, T. Bachteler, and J. Reiher. Privacy-preserving record linkage using bloom filters. *BMC Medical Informatics and Decision Making*, 2009.

[18] W. Winkler. Overview of record linkage and current research directions. Technical report, 2006.

[19] X. Yang, B. Wang, and C. Li. Cost-based variable-length-gram selection for string collections to support approximate queries efficiently. In *In SIGMOD Conference*, pages 353–364, 2008.

[20] A. C.-C. Yao. How to generate and exchange secrets. In *Foundations of Computer Science, 1986., 27th Annual Symposium on*, 1986.