Privacy-Preserving Data Linkage and Geocoding: Current Approaches and Research Directions

Peter Christen
Department of Computer Science, The Australian National University
Canberra ACT 0200, Australia
Peter.Christen@anu.edu.au

Abstract

Data linkage is the task of matching and aggregating records that relate to the same entity from one or more data sets. A related technique is geocoding, the matching of addresses to their geographic locations. As data linkage is often based on personal information (like names and addresses), privacy and confidentiality are of paramount importance. In this paper we present an overview of current approaches to privacy-preserving data linkage, and discuss their limitations. Using real-world scenarios we illustrate the significance of developing improved techniques for automated, large scale and distributed privacy-preserving linking and geocoding. We then discuss four core research areas that need to be addressed in order to make linking and geocoding of large confidential data collections feasible.

1. Introduction

Many businesses and organisations are collecting, storing, processing, analysing and mining fast-growing sets of data containing tens or even hundreds of millions of records. Often this data is about people and contains names, addresses, dates of birth, and other personal information. Analysing and mining such data frequently requires multiple data sources to be combined, linked and aggregated in order to enable more detailed analysis, and allow studies that otherwise would have been impossible. *Data* or *record linkage* [4, 16] has traditionally been used in statistics and in the health sector. Nowadays, data linkage is increasingly being applied in many businesses, as well as in and between government agencies to improve outcomes in taxation, census, immigration, social welfare, in crime and fraud detection, and in the assembly of terrorism intelligence.

A technique related to data linkage is geocoding [5], the matching or linking of addresses to a reference database of standardised and validated addresses and their geographic locations. Geocoding is a significant initial step before data can be loaded into geographical information systems, and before it can be spatially analysed, mined or visualised.

The data linkage process is often challenged by the lack of common unique entity identifiers [16]. Thus, the available common attributes (like person identifiers, addresses, and other data specific information) have to be used in the linkage process. These attributes, however, can contain typographical errors, they can be coded differently, parts can be out-of-date or swapped, or even be missing. In the classical probabilistic approach [16], pairs of records from two data sets are compared and classified as *matches*, *non-matches*, or as *possible matches* (those record pairs for which manual *clerical review* is needed to decide their final linkage status). In recent years, researchers have started to explore the use of techniques from machine learning, data mining, information retrieval, and artificial intelligence to improve the linkage process [4, 16]. Many of these new approaches are based on supervised learning techniques and require training data, which is often not available in real world situations, or only obtainable via manual preparation (a costly process similar to manual clerical review).

When linking two data sets, each record in one data set potentially has to be compared with all records in the second data set. The possible number of record pairs equals the product of the sizes of the two data sets. Techniques known as *blocking* [4] are applied to reduce the number of record pairs that will be compared. They cluster records into blocks, and only compare records within the same block.

Linking or geocoding today's massive data sets with millions of records has the following three major challenges.

1. Even when applying blocking, the computational requirements result in very large memory foot-prints and in long run-times even on powerful modern machines.
2. Comparing a very large number of record pairs will result in many pairs being classified as possible matches, and the manual clerical review process therefore becomes more time consuming, or even impossible. To-

tal project times of several weeks for large linkages using current techniques and involving several linkage experts are not uncommon.

3. Another major challenge in data linkage and geocoding are privacy and confidentiality concerns that arise when personal or confidential data is used for linking. Protecting the personal details of individuals is paramount. Widespread application of data linkage will only gain public acceptance if privacy and confidentiality of such data collections are guaranteed.

New computational techniques are required for increased linkage performance on modern parallel and distributed computing platforms, and automated decision models are needed that will reduce or even eliminate the manual clerical review step while keeping a high linkage quality. Privacy-preserving linking and geocoding techniques are required to allow the linking of data collections between organisations without revealing any personal or confidential information. While partial solutions exists to all three challenges, to the best of our knowledge no currently available linkage technique is tackling all three.

The contributions of this paper are to provide an overview of the currently available privacy-preserving data linkage techniques and to identify four core research areas that need to be addressed in order to make the automated and distributed privacy-preserving data linkage and geocoding of very large data collections possible.

2 Data linkage and geocoding scenarios

In the following we illustrate privacy and confidentiality issues arising from data linkage and geocoding through several scenarios taken from real world situations.

Scenario 1: *An epidemiologist is interested in analysing the effects of car accidents upon hospital admissions, for example what types of injuries are most common, the resulting financial burden upon the public health system, or the general health of people involved in serious car accidents. To be able to conduct such an analysis, the researcher needs access to hospital data, as well as detailed data from car insurers and possibly even access to a police database.* □

In this scenario, the researcher might be able to get access to all source data containing identifying information. Alternatively, the data could be transfered to a trusted proxy organisation (for example a linkage unit within a government health department), which performs the linkage and only provides the linked data (without identifying information) to the researcher. In both cases the original data has to be made accessible to the party conducting the linkage. This might prevent an organisation from being able or willing to participate towards such a valuable project.

Scenario 2: *A population based cancer register aims to geocode its data in order to conduct a spatial analysis of different types of cancer in its region. Due to limited resources the register cannot invest in an in-house geocoding system (i.e. software and personnel) but is reliant on an external geocoding service.* □

The legal or regulatory framework might not allow the cancer register to send their data to an external organisation for geocoding. Even if allowed, complete trust is needed in the capabilities of the external geocoding service to conduct accurate geocode matching, and to properly destroy the register's address data afterwards.

Scenario 3: *Two pharmaceutical companies are interested in collaborating on the expensive development of new drugs. Initially, the companies wish to identify how much overlap of confidential research data there is in their databases (in order to determine the viability of the proposed collaboration), but without having to reveal any confidential data to each other.* □

This scenario requires techniques that allow sharing of large amounts of data in such a way that similar data items are found (and revealed to both companies) while all other data is kept confidential. The involvement of a third party to undertake the linkage will be undesirable to both companies due to the risk of collusion of the third party with either company, or potential security breaches at the linkage party by intruders or its own staff.

Scenario 4: *A honest but curious researcher has access to linked data sets that were provided to the researcher's organisation over a period of time through several research projects. While the linked data sets separately do not contain details that allow identification of individuals, the researcher is able to match records from a midwives data set with records from a HIV database using the commonly available attributes (like postcode, and year and month of birth of mothers). Using a public Web site containing birth notifications, the researcher is able to positively identify births in regional areas by mothers whose details are stored in the HIV database, as year and month of birth of babies are also available in the midwives data set.* □

This scenario highlights the need for techniques that prevent re-identification through linking of several data sets (including publicly available data), that individually only contain de-identified data.

As illustrated by these scenarios, secure techniques are needed that allow the efficient linking and geocoding of large data sets without any possibility that personal or confidential information can leak or be compromised.

3. Current approaches

Traditionally, data linkage techniques have required that all the identifying data is revealed to the linkage party (often a third party like a research group or their proxy). Good practice dictates that medical and other substantive attributes are removed from the records before passing them to the linkage party [8]. This, however, does little to obfuscate the source of those records. Furthermore, the linkage party will gain access to all records in all the data sets to be linked, because there is no way of knowing prospectively which records will match. Thus, traditional data linkage methods require the disclosure of confidential information about large numbers of individuals (albeit to a small number of people who actually undertake the linkage), which clearly invades the privacy of all individuals concerned, and requires complete trust in the intentions of the parties involved, and their ability to maintain confidentiality, as well as security of their computing and networking systems.

However, the invasion of privacy could be avoided, or at least mitigated, if there were some method of determining which records in two data sets matched, or were likely to match on more detailed comparison, without either data source having to reveal any identifying information to each other or to a third party. De-identified versions of the linked records can then be used for subsequent analysis.

First methods based on cryptographic techniques that implement this idea were proposed by a team of French researchers [7]. These methods, which use keyed one-way hash encoding functions, allow the party undertaking the linkage to use all of the identifying data items available in the data sets to be linked, but without the linkage party seeing any of the actual values of those data items. Unlike traditional data linkage techniques, these methods provide good protection against a single party, acting alone, attempting to invade privacy or breach confidentiality. Distributed secure data linkage using keyed one-way hash encoding functions has subsequently been described in [12]. However, this work is limited to exact matching only and does not address the important issue of typographical and other errors which occur in most real world databases.

A three-party protocol termed *blindfolded record linkage* based on q-grams is presented in [6]. It allows for approximate matching by calculating the Dice co-efficient similarity measure between hash-encoded sets of q-grams. The computational and communication overheads of encoding q-gram sets make the approach currently impractical for linking large data sets, or data containing long sequences such as those used in genomics [6]. Two similar protocols for data linkage and cohort extraction (without revealing the membership of any individual in the cohort to the data source) are presented in [10]. They are also based on hash encoded values and improve the security weaknesses of [6].

A secure two-party protocol for string distances, including TF-IDF (commonly used in information retrieval) and the Euclidean distance is discussed in [11]. This protocol is based on a stochastic scalar product. Another two-party protocol for secure and private sequence comparisons based on the commonly used edit-distance approach is presented in [2]. It applies homomorphic encryption in such a way that neither party at any time has information about the complete dynamic-programming matrix used for the edit-distance calculation (as this would allow one party to infer details about the original data held by the other party).

One crucial issue when linking large data sets is blocking, the techniques applied to reduce the number of record pair comparisons [4]. A first set of methods for privacy-preserving blocking has recently been presented in [1]. A secure three-party protocol based on hash encoded values and TF-IDF (similar to [11]) is used, and three different blocking methods are discussed. The basic idea is to compare records only if they have at least one token (e.g. a word) in common (hash encoded binary representations of the records are used). Security issues are discussed and experimental results using smaller data sets (with around 5,000 records each) are presented, showing the practicality of the approach. To our knowledge, this is the only work that so far has been done in this area.

We are not aware of any research specific to privacy-preserving geocoding. While similar to data linkage, geocode matching [5] is specific in that user addresses are linked with a large database of cleaned and standardised reference addresses, and approximate matches have to be handled in special ways. For example, if a given street number in a user address is not available in the reference data, the location of this address should be extrapolated using reference addresses from the same street. Similarly, if an address cannot be found in its given postcode or suburb area, the matching system should extended its search to neighbouring areas [5]. A privacy-preserving geocoding approach should allow an organisation to locally encode their address data and transfer them to a geocoding service, without having to reveal any of these addresses, and without the geocoding service learning which addresses have been matched.

Besides the work done in privacy-preserving data linkage, there is also intense interest in the knowledge discovery and database communities in *privacy-enhanced* data mining [14] and *secure multi-party computation*, as well as *secure information sharing*. Although almost any function can be computed securely without revealing its inputs, all of the presented protocols do so at the expense of communication and computational overheads.

To summarise, many of the presented approaches to privacy-preserving data linkage are currently in an proof-of-concept or prototype state, in that they have been eval-

uated on only small data sets, while other approaches are limited to exact matching only. Many cryptographic techniques have computational and communication overheads that make the linkage of very large data set currently not feasible. Additionally, none of the privacy-preserving techniques has been investigated with the use of machine learning based automated record pair classification in mind.

4. Research directions

To the best of our knowledge, no work into the overall development of large-scale automated and distributed privacy-preserving data linkage and geocoding has so far been conducted. In the following we discuss the four core research areas that have to be addressed to achieve this overall goal.

4.1 Improved secure matching

In the previous section we have presented various approaches based on cryptographic protocols using either two- or three-party protocols. Some of these methods offer only partial privacy protection [8] or restrict the way linkage can be performed [7], while other methods are limited to exact matching only [12]. Only in the last three years have methods been developed that allow approximate matching without the need of the original values being revealed to other parties [1, 2, 6, 11]. These methods compute secure functions at the expense of communication and computational overheads. However, they are partial solutions, in that they don't allow the fully automated linking or geocoding of very large data sets, neither using the traditional probabilistic linkage approach, nor using one of the recently developed machine learning based techniques [4, 16].

Research in this area should aim to develop frameworks that allow the inclusion of a wide variety of secure approximate string comparisons techniques, including the commonly used Jaro and Winkler comparators [16], which so far have not been converted into a privacy-preserving setting [6]. Secure similarity comparison techniques for numerical, date, age, as well as more complex structured data values should also be investigated. All these techniques have to be considered in combination with privacy-preserving blocking [1], so that linking of very large data sets becomes feasible. It is also important to develop new methods for privacy-preserving linkage that have reduced communication and computational overheads compared to current methods, as otherwise linking very large data sets will be problematic. Developing protocols specific for privacy-preserving geocoding will also be of importance, in order to facilitate applications that allow organisations to geocode their addresses without having to reveal them to any other organisation.

4.2 Automated record pair classification

This second area of research is important as it will leverage the methods developed in the first area, allowing automated data linkage and geocoding without human intervention. None of the linkage methods based on machine learning, artificial intelligence and information retrieval techniques developed in the past few years [16] take privacy preservation into account. Many are using supervised learning techniques, and thus require training data that often has to be prepared manually. As within a privacy-preserving setting only encoded data is available to the linkage party, neither supervised learning nor the traditional manual clerical review process (for possible matches) are feasible.

Research in this area therefore has to concentrate on the development of unsupervised secure classification techniques. While initial work on clustering and hierarchical graphical models have shown to be promising in the context of data linkage, no work has so far been done in using such techniques within a secure setting. Unsupervised techniques have to be reconsidered from a privacy preservation point of view. Techniques being developed in privacy-preserving data mining [14] will have to be modified in order to become suitable for data linkage applications.

Enabling automatic linking and geocoding in a privacy-preserving setting will significantly impact on the productivity of the organisations undertaking such linkages, as it will free up the human resources currently needed for the tedious manual clerical review process or the manual preparation of training examples.

4.3 Scalability

While secure matching and automated classification techniques are at the core of privacy-preserving data linkage, computational requirements still challenge the linking and geocoding of very large data sets with tens or even hundreds of millions of records. Techniques need to be developed that allow distributed linking and geocoding on modern computing environments like parallel and high-performance computers, clusters and computational grids.

Only limited research has so far been done in this area. Work in [3] showed that parallel data linkage can achieve good speedup results, as the computationally expensive comparison of record pairs can be done with only little communication overhead, assuming all data is available on all computing nodes. This assumption, however, will not hold for parallel and distributed platforms like clusters or computational grids, or when linkage is done between different organisations, possibly using a third party to perform the linkage. Different parallelisation approaches need to be developed to achieve scalability both with the size of the data sets and the number of computing nodes used.

Being able to securely link very large data sets in short time periods will significantly improve the productivity of a linkage organisation and result in faster delivery of the linked data to the end-user. In scenarios like an outbreak of a highly contagious disease or a suspected (bio-) terrorism attack it is absolutely crucial to get linkage or geocoding results in near real-time (seconds or minutes).

4.4 Preventing re-identification

While this research area is outside the core data linkage and geocoding functionality, it is nevertheless very important and has to be considered carefully, as otherwise all efforts made in privacy-preserving linking can become useless. As shown in Scenario 4 in Section 2, while properly de-identified linked data by itself does not allow re-identification, if linked to other data (possible from earlier linkages or publicly available) it can become feasible to re-identify certain records. This can obviously result in a loss of privacy and confidentiality for the individuals whose records are being re-identified.

A large body of work has been done in statistics on micro-data confidentiality [15]. This includes techniques for masking data (like swapping or aggregating values) so that it can be made public while reducing the risk of re-identification. Research done in the security and data mining communities, for example on k-anonymity [13] and trail re-identification [9], is also highly relevant. Such approaches will have to be investigated further, with the aim to fully integrate them into privacy-preserving data linkage and geocoding systems, so that during the linkage process information about potential re-identification can be collected, identified and dealt with.

5 Conclusions

We have presented an overview and discussed the limitations of current approaches to privacy-preserving data linkage and geocoding, and we have outlined four core research areas that need to be addressed in order to make large scale and distributed privacy-preserving data linkage and geocoding practical. Techniques from cryptography, data mining, machine learning, and high-performance and distributed computing will have to be synthesised to develop a new generation of secure, automated, efficient and accurate techniques for linking and geocoding of very large data sets with millions of records.

Acknowledgements

This work is supported by an Australian Research Council (ARC) Linkage Grant LP0453463 and partially funded by the New South Wales Department of Health.

References

[1] A. Al-Lawati, D. Lee, and P. McDaniel. Blocking-aware private record linkage. In *IQIS '05: Proceedings of the 2nd international workshop on Information quality in information systems*, pages 59–68, Baltimore, 2005. ACM Press.

[2] M. J. Atallah, F. Kerschbaum, and W. Du. Secure and private sequence comparisons. In *WPES'03: Proceedings of the 2003 ACM workshop on Privacy in the electronic society*, pages 39–44, Washington DC, 2003. ACM Press.

[3] P. Christen, T. Churches, and M. Hegland. Febrl – a parallel open source data linkage system. In *PAKDD, Springer LNAI 3056*, pages 638–647, Sydney, 2004.

[4] P. Christen and K. Goiser. Quality and complexity measures for data linkage and deduplication. In F. Guillet and H. J. Hamilton, editors, *Quality Measures in Data Mining*, volume 43 of *Studies in Computational Intelligence*. Springer, 2006.

[5] P. Christen, A. Willmore, and T. Churches. A probabilistic geocoding system utilising a parcel based address file. In *AusDM'04*, pages 130–145, Cairns, Australia, 2006. Springer LNAI 3755.

[6] T. Churches and P. Christen. Some methods for blindfolded record linkage. *BioMed Central Medical Informatics and Decision Making*, 4(9), 2002.

[7] L. Dusserre, C. Quantin, and H. Bouzelat. A one way public key cryptosystem for the linkage of nominal files in epidemiological studies. *Medinfo*, 8(644-7), 1995.

[8] C. W. Kelman, J. A. Bass, and D. Holman. Research use of linked health data – a best practice protocol. *ANZ Journal of Public Health*, 26(3), 2002.

[9] B. Malin and L. Sweeney. A secure protocol to distribute unlinkable health data. In *American Medical Informatics Association 2005 Annual Symposium*, pages 485–489, Washington DC, 2005.

[10] C. M. O'Keefe, M. Yung, L. Gu, and R. Baxter. Privacy-preserving data linkage protocols. In *WPES'04: Proceedings of the 2004 ACM workshop on Privacy in the electronic society*, pages 94–102, Washington DC, 2004.

[11] P. Ravikumar, W. W. Cohen, and S. E. Fienberg. A secure protocol for computing string distance metrics. In *PSDM held at ICDM*, Brighton, UK, 2004.

[12] G. Schadow, S. J. Grannis, and C. J. McDonald. Discussion paper: privacy-preserving distributed queries for a clinical case research network. In *CRPIT '14: Proceedings of the IEEE international conference on privacy, security and data mining*, pages 55–65, 2002.

[13] L. Sweeney. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzz. Knowl.-Based Sys.*, 10(5), 2002.

[14] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, S. Yucel, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *SIGMOD Rec.*, 33(1):50–57, 2004.

[15] W. E. Winkler. Masking and re-identification methods for public-use microdata: Overview and research problems. In *Privacy in Statistical Databases*, pages 216–230, Barcelona, 2004. Springer LNCS 3050.

[16] W. E. Winkler. Overview of record linkage and current research directions. Technical Report RRS2006/02, US Bureau of the Census, 2006.