Secure Health Data Linkage and Geocoding: Current Approaches and Research Directions

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

Tim Churches
Centre for Epidemiology and Research,
New South Wales Department of Health,
Locked Mail Bag 961, North Sydney NSW 2059
tchur@doh.health.nsw.gov.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. In the health sector, data linkage is commonly used to assemble longitudinal or epidemiological data sets that would otherwise not be available, and geocoding is employed for spatial analysis of health data. As data linkage is often based on personal information (like names, addresses, and dates of birth), privacy and confidentiality issues are of paramount importance.

In this paper we present an overview of current approaches to secure data linkage and geocoding and discuss their limitations, and using several real-world scenarios we illustrate the significance of developing improved techniques for large scale and distributed secure linking and geocoding. We discuss four core areas of research that need to be addressed in order to make linking and geocoding of large confidential data collections possible: secure matching techniques, automated record pair classification, scalability, and techniques that prevent re-identification of records over collections of linked data. Finally, we give a short overview of several Australian projects in this area.

Keywords: record linkage, privacy preservation, geocode matching, cryptography.

1 Introduction

Many organisations in the health sector are collecting, storing, processing and analysing increasingly large data collections with millions of records. Most of this data is about patients and contains identifying (such as names, addresses, and dates of birth), as well as confidential information (such as details of medical procedures and tests). Analysing such data often requires information from multiple sources to be linked and aggregated in order to enable more detailed analysis, and allow studies that otherwise would have been impossible. Today, health data linkage not only faces computational and operational challenges due to the increasing size of data collections and their complexity, but also privacy and confidentiality challenges due to growing concerns by the general public about their personal information being linked and shared within and between health organisations [8, 11, 18].

Data or *record linkage* (also known as *data matching*, *data integration*, or *data cleaning*) has traditionally been used in statistics for linking census data [21] and in the health sector for longitudinal and epidemiological studies [7, 11]. Today, data linkage techniques are increasingly being applied in and between government organisations to improve outcomes in taxation, census, immigration, social welfare, in crime and fraud detection, and in the assembly of terrorism intelligence [18].

*Corresponding author

A technique related to data linkage is geocoding [6], the matching or linking of addresses (that can contain typographical and other errors, be incomplete, or out-of-date) to a reference database of standardised and validated addresses and their geographic locations (latitude and longitude). Geocoding is significant, as it is the initial step before data can be loaded into geographical information systems, and before it can be spatially analysed and visualised. Spatial data analysis is crucial, for example when dealing with outbreaks of rapidly spreading contagious diseases, or when investigating (bio-) terrorism intelligence. Accurate linkage of addresses is important, as any subsequent data processing, visualisation and analysis depends upon the quality of the linked data.

Computer-assisted data linkage goes back as far as the 1950s, and the mathematical foundation of probabilistic data linkage (as developed by Fellegi and Sunter in 1969) is still the basis of many current linkage systems [21]. Often the linkage process is challenged by the lack of a common unique entity identifier, and thus becomes non-trivial. In such cases, person identifiers (like names and dates of birth), demographic information (like addresses) and other specific information (like medical details) have to be used to achieve good linkage results. 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 [21], pairs of records from two data sets are compared using various similarity functions (like exact or approximate string, numerical, date, or age comparisons) and then classified into *matches* (if the compared attributes mainly agree), *non-matches* (if the compared attributes mainly disagree), or as *possible matches* (if the linkage system cannot make a clear decision). The class of possible matches are those record pairs for which manual *clerical review* is needed to decide their final linkage status. Data linkage of two data sets $\mathbf{A}$ and $\mathbf{B}$ considers record pairs in the product space $\mathbf{A} \times \mathbf{B}$ and determines which pairs are matches. Thus, the total number of record pairs equals the product of the sizes of the two data sets, i.e. $|\mathbf{A}| \times |\mathbf{B}|$, where $|\cdot|$ denotes the number of records in a data set. Comparing all pairs is computationally only feasible for small data sets containing up to several thousand records each, as, for example, linking two data sets with $100,000$ records each would result in 10^{10} (ten billion) record pair comparisons. Techniques known as *blocking* [5, 21] are applied to reduce the number of record pair comparisons. They cluster records into blocks and only compare records within the same block, thereby reducing the complexity of the overall linkage process.

In recent years, computer science researchers have started to explore the use of various techniques taken from machine learning, data mining, database research, information retrieval, and artificial intelligence to improve the linkage process [5, 21]. Techniques investigated include learning the optimal parameters for approximate string comparison techniques (like edit-distance costs); representing records as document vectors (an approach taken from information retrieval); applying active learning (a technique where the learning system selects difficult pairs of records for manual classification, thereby reducing human intervention); using supervised learning approaches (where manually prepared training data, i.e. pairs of classified records, are needed to train a classifier); and clustering (unsupervised learning techniques that explore the structure of the data without the need of manual training examples). Many of these new approaches, however, do 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). Additionally, many of the recent publications in this area present experimental linkage studies that are based on only small data sets with a couple of thousand records.

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

1. Even when using blocking the computational requirements (memory usage and processing time) result in linkage run-times of hours even on powerful modern machines. For example, linking two data sets with $5,000,000$ records each and a blocking technique that reduces the number of record pairs from 2.5×10^{13} to $100,000,000$ (so that in average each record in one data set is compared to twenty records in the other data set), assuming that $10,000$ record pairs can be compared per second (0.1 milli-second per comparison), will take almost three hours.

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. For the above example, if only 0.1% of the compared record pairs are classified as possible matches, manual review is required for 100,000 record pairs. This will be a very tedious task requiring expensive human resources. Total project times of several weeks for large linkages using current techniques and involving several linkage experts are not uncommon.

3. The third 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, especially in the health sector, where a breach of privacy can result in a person's medical history being compromised. New application areas of data linkage (like electronic health records stored on smart-cards that can be accessed by doctors, public and private health insurers, as well as the national health administration system) will only gain public acceptance if privacy and confidentiality of all records in 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. Secure (also named *privacy-preserving*) linking and geocoding techniques are required to allow the linking of large data collections within and between health organisations without revealing any personal or confidential information. While partial solutions exists to all three of the above challenges, to the best of our knowledge no currently available linkage approach is tackling all.

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

2 Data linkage and geocoding scenarios

While analysing linked or geocoded data can be beneficial in areas like health and crime and terror detection, many individuals are increasingly worried about their personal information being collected, linked and shared by various organisations. Linking and geocoding data can result in a breach of privacy for the individuals involved, or a loss of confidential information for an organisation, resulting in the rejection of data linkage and geocoding by the general public as well as private and public organisations. In the following we illustrate these issues using several heath linkage scenarios.

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, and the general health of people that were involved in serious car accidents. To be able to achieve 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.* $\diamondsuit$

In this scenario, the researcher might be able to get access to all source data containing identifying information (following proper regulatory procedures, like getting approval from ethics committees, signing confidentiality agreements, etc.), in which case the linkage can be performed by the researcher (or a support entity at the researcher's university) following strict security and access limitations. 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, however, the original data (encrypted only for transfers between organisations) has to be made available to the party undertaking the linkage (i.e. the original unencrypted identifying values are needed for the linkage). This limitation might prevent an organisation from being able or willing to provide their data towards such a linkage project, and thus prevent an analysis that would be of significant benefit.

Scenario 2: *A population based cancer register aims to geocode its data in order to conduct a spatial analysis of different types of cancer. 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 organisation to conduct accurate geocoding, and to properly destroy the register's address data afterwards. If the geocoding organisation is a commercial company, limited independent information will be available to the register about its matching performance. Alternatively, the register might be able to use the geocoding service of a trusted proxy organisation, like a government health department. In both cases, the original addresses have to be made available to the outside organisation that performs the geocoding.

Scenario 3: *Two pharmaceutical companies are interested in collaborating on the expensive development of new drugs. Before initiating the collaboration the companies wish to identify how much overlap of confidential research data there is in their databases (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 information is kept confidential. Such techniques would thus prohibit any data from one company being available in its original form to the other company, and vice versa. The involvement of a third party to undertake the linkage will be undesirable to both companies due to the risk of collusion by the third party with either company, or potential security breaches at the third party by intruders.

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 allow identification of individuals, the researcher is able to match records in a midwives data set with records in 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, possibly including data that is publicly available, that individually only contain de-identified data (i.e. data that does not allow re-identification).

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. In the following section we present partial solutions that have been developed to tackle this challenge, and in Section 4 we discuss four core research areas needed to make large scale distributed secure data linkage and geocoding possible.

3 Current approaches

Traditionally, data linkage techniques have required that all the identifying data in which links are sought be revealed to at least one party, often a third party (for example the researchers or their proxy). Good practice dictates that medical and other substantive attributes should be removed from the records before passing them to a person or organisation undertaking the data linkage operation [7, 11]. This, however, does little to obfuscate the source of those records. In many circumstances knowledge of the data source permits significant and highly confidential information to be inferred about individuals who are identified in the candidate records to be linked. Furthermore, the party undertaking the linkage necessarily requires access to all records in all the data sets to be linked, because there is no way of knowing prospectively which records will match. Traditional data linkage methods thus

require the disclosure of confidential information about large numbers of individuals, albeit to a small number of people who actually undertake the linkage. This approach 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, and security of their computing and networking systems. It is typically infeasible to obtain consent for this invasion of privacy from all individuals identified in each of the databases, instead one or more ethics committees or institutional review boards must consent for the linkage on behalf of all the individuals involved.

Various approaches and protocols on how to better protect the privacy of individuals whose records are to be linked have been developed in recent years, mainly in the health sector. One approach is to severely limit the identifying data items which are disclosed to the data linking party. This approach has the disadvantage that as the number and details of the (partially) identifying data items which are disclosed to the linkage party are reduced, the accuracy and overall efficiency of the linkage operation are diminished. Another approach is to physically separate the identifying attributes from medical or other sensitive data and to use a highly trusted third party to undertake the linkage. A protocol based on this idea is described in [11], and a variation of this approach is currently being used by the *Western Australian Data Linkage Unit*.[1] A similar approach aimed at population based disease registers is discussed in [7], where medical details are separated from person identifiers and encrypted using different keys, and a trusted third party is responsible for obfuscating the sources of records before sending them to a single population register that is responsible for linking personal details and maintaining unique person identifiers.

The invasion of privacy could be avoided, or at least mitigated, if there were some method of determining which records in two databases matched, or were likely to match on more detailed comparison, without either database 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 analysis.

First methods based on cryptographic techniques that implement this idea were proposed in the mid 1990s by French researchers [9]. These methods, which use keyed one-way hash encoding functions [16], allow the party undertaking the linkage to use the partially-identifying data items available in the data sets to be linked, but without this 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 [15]; however, this work does not address the important issue of typographical errors and other variations which occur in most real data (the method is limited to exact matching only).

In general, cryptographic approaches to secure data linkage and data sharing [12] can be classified into two- and three-party protocols, as illustrated in Figure 1, and discussed below.

3.1 Secure two-party protocols

In a two-party protocol, the two data sources, named *Alice* and *Bob*, wish to share or link data in such a way that only information about the shared data is revealed to both parties. The general approach of two-party protocols consists of the following three steps.

(1) The two parties agree on a secret random key, which they share only with each other. The *Diffie-Hellman* key agreement protocol [16] can be used for this. All subsequent transfers of data between the two parties are assumed to be authentic and secure through the use of a public key infrastructure (PKI) [16] (the agreed secret key is used to sign and encrypt all messages).

(2) Both parties pre-process, transform and encode their data according to an agreed manner (they might also add *chaff* to their data in the form of dummy records to obfuscate the original records). Each party then sends this encoded data to the other party.

[1] URL: `http://www.dla.org.au`

Figure 1: Basic two- and three-party linkage protocols.

(3) Each party now performs the linkage using their own and the encoded data received from the other party, and then returns information about the linked records back to the other party. This information might only be the number of similar records in common in both parties, or the identifiers of these records. Depending upon the information exchanged, a party might be able to infer more details about the other party's data.

Depending upon the actual linkage technique employed, steps (2) and (3) might be repeated several times. A most important requirement for any two-party protocol is that at any time no party will have all the information needed to infer the original record values held by the other party.

A secure two-party protocol for string distances is discussed in [14]. This protocol is based on a stochastic scalar product, that is provably as secure as the underlying set-intersection cryptographic protocol it is using. Another two-party protocol for secure sequence comparisons based on the edit-distance approach is presented in [3] (edit-distance is calculated as the minimum number of character inserts, deletes and substitutions needed to transform one string sequence into the other). It applies encoding 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 a party to infer details about the original data held by the other party).

3.2 Secure three-party protocols

Three- or third-party protocols for secure data linkage are based on the idea that a (more or less trusted) third party, *Carol*, performs the linkage, without either of the two data sources having to reveal any identifying information to any other party. Similar to two-party protocols, the general approach of three-party protocols consists of the following three steps.

(1) The two data sources, Alice and Bob, again mutually agree on a secret random key, which they share only with each other, but not with the linkage party Carol.

(2) Both parties pre-process, transform and encode their data according to an agreed manner and using the secret key, and then send the encoded data to the linkage party Carol, which performs the linkage without seeing any of the original values. It is assumed that the communication between the two data sources and the linkage party is secured using PKI with two different keys used between Alice and Carol and Bob and Carol.

(3) Information about the linked data is sent from Carol to both data sources. Again, this might only be the number of similar records in common, or include the identifiers of these records.

Several three-party protocols for privacy-preserving data linkage have been developed in the last few years, with different techniques of how the linkage party is calculating the similarity between values, and with different amounts of information that can be inferred by any of the parties.

A protocol termed *blindfolded record linkage* based on q-grams is presented in [8] (q-grams are sub-strings of length q, for example, 'peter' contains the 2-grams 'pe','et','te', and 'er'). The protocol allows for approximate matching of values with typographical errors by calculating the Dice co-efficient

similarity measure between hash-encoded sets of q-grams. Weaknesses of this protocol include that the third party, Carol, could mount a frequency analysis attack against the encrypted q-gram sets and compare them to frequencies in similar data (like names taken from a telephone directory), while a second threat is that Carol is colluding with Alice (or Bob) in an attempt to discover Bob's (or Alice's) values. Several remedies are described [8], including the last minute election of the linkage party from a collection of many functionally equivalent parties. The computational and communication overheads of encoded q-gram sets make the approach currently impractical for linking very large data sets, or data containing long sequences such as those used in genomics or proteomics.

Two three-party protocols for data linkage and cohort extraction (without revealing the membership of any individual in the cohort to the data source) are presented in [13]. They are based on hash encoded values and improve the security weaknesses of [8]. Building on ideas presented in [1] on information sharing in private databases, the two protocols put together allow a third party (e.g. a researcher requiring access to the linked data) to construct a linked data set so that (1) no identifying information is revealed to any other party by any data source, and (2) no data source learns which data has been extracted from their database. The presented protocols have good security characteristics and minimise information leakage.

3.3 Secure blocking

One crucial issue when linking large data sets is blocking, the techniques applied to reduce the number of record pair comparisons. First methods for secure blocking have recently been presented in [2]. A secure three-party protocol based on hash encoded values and string distance calculations (similar as done in [14]) 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 encrypted binary representations of the records are used). Security issues are discussed and experimental results using smaller data sets 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.

3.4 Secure geocoding

Secure data linkage techniques can also be useful for geocode matching. The two basic approaches to geocoding are: (1) the data to be geocoded is sent from the data source to a geocoding service (thereby compromising the privacy of all addresses in the data), and (2) the data source purchases geocoding software and reference data (and then performs geocoding in-house). The advantage of the second approach is that no addresses have to be given to an external organisation; however, the disadvantages are the costs of purchasing a geocoding system and reference data, as well as training of geocoding expertise in-house. This approach will thus only be viable for large organisations.

We are not aware of any research specific to secure geocoding. While similar to data linkage, geocode matching [6] 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 [6].

A secure 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. Several of the approaches presented above can be used for such a task; however, so far only in [8] have some initial ideas based on multiple linkage parties been discussed.

3.5 Secure multi-party computation

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* and *secure multi-party computation*, as well as *secure information sharing* [1, 19]. Although it appears that almost any function can be computed securely without revealing its inputs, all of the presented techniques do so at the expense of communication and computational overheads.

To summarise this overview, many of the presented approaches to secure or privacy-preserving data linkage are currently in an initial proof-of-concept or prototype state, in that they have been evaluated on only small or medium sized data sets (containing some several thousand records), 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 secure techniques has investigated the use of machine learning based automated record pair classification as discussed in Section 1.

4 Research directions

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

4.1 Improved secure matching techniques

Of all four areas this is the one where most research has been done so far. 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 or restrict the way linkage can be performed, while others only allow exact matching. Only in the last three years have methods been developed that allow approximate matching of strings without the need of the original values being revealed to other parties [2, 3, 8, 14]. 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 one of the recently developed approaches discussed in Section 1.

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-Winkler comparator [21], which so far has not been converted into a privacy-preserving setting [8]. Secure similarity comparison techniques for numerical, date, age, as well as more complex structured data values should be investigated as well.

It is also important to develop new methods for secure linkage that have reduced communication and computational overheads compared to current methods, as otherwise linking of very large data sets will be problematic. Secure approaches for both two- and three party protocols are needed for a variety of similarity comparison techniques in order to facilitate privacy-preserving linkage and geocoding of data sets with various characteristics and contents in different scenarios. Additionally, all these techniques have to be considered in combination with secure blocking [2] so that linking of very large data sets becomes feasible. Modifying the developed protocols and methods so that secure geocoding can be performed will also be of importance.

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. Many linkage meth-

ods based on machine learning, artificial intelligence and information retrieval techniques have been developed in the past few years. However, none of these methods takes security and privacy preservation into account. Most are based on 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 party undertaking the linkage, neither supervised learning nor the traditional clerical review process of manually classifying 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 developed in privacy-preserving data mining [19] and machine learning will have to be modified in order to become suitable for secure 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 secure 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.

Being able to securely link large data sets in short time periods will significantly improve the productivity of the party undertaking the linkage 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.

Only limited research has so far been done in this area [4, 10, 15]. Some recent work has shown that parallel data linkage can achieve good speedup results [4], as the computationally expensive comparison of record pairs can be done with only little communication, assuming all data is available on all computing nodes (like on a shared memory multiprocessor). This assumption, however, will not hold for platforms like clusters or grids, or when linkage is done between organisations, possibly using a third party to perform the linking. Different parallelisation approaches have to be developed to achieve scalability both with the size of the data sets to be linked and the number of processors computing nodes used.

Computational issues that need to be considered in heterogeneous distributed computing environments include data distribution and load balancing (due to potentially dynamically changing loads on the computing nodes used), fault tolerance (due to interrupted network connections and node failures), as well as scalability (the question of how many nodes to use for a given linkage or geocoding problem), and the optimal ratio of communication to computation for a given environment (which might change dynamically at runtime). Addressing these questions within the framework of secure data linkage will result in practical techniques for linking and geocoding large data sets. Additionally, issues like access and charging policies for data linkage and geocoding services, as well as having suitable user interfaces, have to be solved as well.

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 [20]. 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 [17], is also highly relevant. Such approaches will have to be investigated further, with the aim to fully integrate them into secure data linkage and geocoding systems, so that during the linkage process information about potential re-identification can be collected, identified and dealt with.

5 Research in Australia

There are several projects currently being conduced in Australia in the area of data linkage and geocoding. The *Western Australian Data Linkage Unit*, based on a best practise protocol [11], is currently re-designing it's linkage infrastructure. In Queensland, the *Health Data Integration (HDI)* project [10], a collaboration between CSIRO and Queensland Health aims at developing infrastructure and software (using secure linkage protocols) specific to the health sector which will allow health professionals to access aggregated shared data while maintaining the privacy of patient data. The *Febrl (Freely extensible biomedical record linkage)* project, a collaboration between the Australian National University and the NSW Department of Health, is developing new and improved techniques for large scale high-performance data linkage [4], geocoding [6], and privacy-preserving data linkage [8]. The project has also developed an open source (freely available) prototype data linkage software.[2]

6 Conclusions

We have presented an overview and discussed the limitations of current approaches to secure health 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 secure 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.

While the four research areas discussed in this paper focus on computational and privacy-preserving technical challenges, a fifth major challenge lies in achieving public acceptance for these techniques, which in turn will allow appropriate legal and regulatory frameworks be put into place. In many countries, including Australia, public perception towards data linkage, and the potential of privacy and confidentiality breaches resulting from linking and geocoding, currently limits the application of these techniques. It is therefore important that new secure techniques for data linkage and geocoding are being discussed and scrutinised by information and network security specialists, health researchers and legal experts, as well as the general public. Only if the advantages of linked data (especially in areas like health, and fraud, crime and terror detection), and the security offered by new secure linkage techniques are becoming accepted by the public, will these techniques become successful.

Acknowledgements

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

[2] URL: `http://datamining.anu.edu.au/linkage.html`

References

[1] R. Agrawal, A. Evfimievski, and R. Srikant. Information sharing across private databases. In *SIGMOD'03*, pages 86–97, San Diego, 2003. ACM Press.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] T. Churches. A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers. *BioMed Central Medical Research Methodology*, 3(1), 2003.

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

[9] 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.

[10] D. Hansen, C. Pang, and A. Maeder. HDI: Integrated services for health data. In *International Conference on Machine Learning and Cybernetics (ICMLC)*, Guangzhou, China, 2005.

[11] 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.

[12] Y. Li, J. Tygar, and J. M. Hellerstein. Private matching. In D. Lee, S. Shieh, and J. Tygar, editors, *Computer Security in the 21st Century*. Springer, 2005.

[13] 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.

[14] 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.

[15] 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.

[16] B. Schneier. *Applied Cryptography: Protocols, Algorithms, and Source Code in C, 2nd edition*. John Wiley & Sons, Inc., New York, 1995.

[17] L. Sweeney. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.

[18] L. Sweeney. Privacy-enhanced linking. *SIGKDD Explorations*, 7(2):72–75, 2005.

[19] 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.

[20] W. E. Winkler. Masking and re-identification methods for public-use microdata: Overview and research problems. Technical Report RRS2004/06, US Bureau of the Census, 2004.

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