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 (latitude and longitude). As data linkage is often based on personal information (like names, dates of birth, and addresses), privacy and confidentiality issues are of paramount importance, especially when linking data across organisations.

In this paper we present an overview of current approaches to privacy-preserving 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 privacy-preserving 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.

1. Introduction

Many organisations are collecting, storing, processing, analysing and mining fast-growing sets of data containing tens or even hundreds of millions of records, with millions of records being added per annum. Examples of such data collections occur in retail, credit card and insurance administration, telecommunication, census, taxation, the health sector, and in security and intelligence agencies. In many cases this data is about people and contains names, addresses, dates of birth, and other personal information. Analysing and mining such large data sets often requires information from multiple data sources to be combined, linked and aggregated in order to enable more detailed analysis, and even allow studies that otherwise would have been impossible [9]. Today, 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 between organisations [18, 31].

Data or *record linkage* (also known as *data matching* or *data integration*) has traditionally been used in statistics for linking census data [49] and in the health sector for longitudinal and epidemiological studies [31]. For example, research in Western Australia based on an ambulance cardiac arrest database linked with hospital data and death registers led to the installation of defibrillators in ambulances and hospital wards, and the appropriate training of nurses as first aid responders, saving many lives [9]. 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 [45]. Many businesses routinely deduplicate and link their data sets when compiling mailing lists, and databases containing customer information are often sold to specialised businesses for marketing purposes.

A technique related to data linkage is geocoding [16], 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, mined or visualised. Spatial data analysis is crucial, for example when dealing with outbreaks of rapidly spreading contagious diseases, or when investigating crime and terrorism intelligence. Accurate linkage of addresses is important, as any subsequent data processing, visualisation, analysis and mining depends upon the quality of the linked data.

Computer-assisted data linkage goes back as far as the 1950s. The mathematical foundation of probabilistic

(or statistical) data linkage as developed by Fellegi and Sunter [26] in 1969 is still the basis of many current linkage systems. Often the linkage process is challenged by the lack of a common unique entity identifier, and thus becomes non-trivial [49]. In such cases, person identifiers (like names and dates of birth), demographic information (like addresses) and other data specific information (like medical details or customer information) 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 [26, 49], pairs of records from two data sets are compared using various similarity functions (like exact or approximate string, numerical, date, or age comparisons). The resulting numerical similarity values for a record pair are summed into a matching weight R. Two thresholds t_{lower} and t_{upper} (with $t_{lower} < t_{upper}$) are then used to classify a record pair:

$$\begin{array}{lll} \text{if} \quad R > t_{upper} & \rightarrow & \textit{match}, \\ \text{if} \quad t_{lower} \leq R \leq t_{upper} & \rightarrow & \textit{possible match}, \\ \text{if} \quad R < t_{lower} & \rightarrow & \textit{non-match}. \end{array}$$

The class of *possible matches* are those record pairs for which human oversight, also known as *clerical review*, is needed to decide their final linkage status. In theory, it is assumed that the person undertaking this review has access to additional data (or may be able to seek it out) which enables her or him to resolve the possible matches. In practice, however, often no additional data is available and the clerical review process becomes one of applying experience, common sense or human intuition to make the decision.

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 of these pairs are matches. The number of possible pairs equals the product of the sizes of the two data sets, so the straight forward approach results in a quadratic complexity of $O(|\mathbf{A}| \times |\mathbf{B}|)$, where $|\cdot|$ denotes the number of records in a data set. This naive approach 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* [4, 15] are applied to reduce the number of record pairs that will be compared. These methods work by clustering records into blocks [22] and only comparing records within the same block, thereby reducing the complexity of the overall linkage process.

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. A popular approach [6, 10, 21, 50, 53] is to learn distance measures (like edit-distance) that are used for approximate string comparisons [13]. As shown in [21], combining different learned string comparison methods can result in improved linkage classification. An information retrieval based approach [20] is to represent records as document vectors and compute the cosine distance between such vectors, while [35] explores the use of support vector machines to classify record pairs. Active learning is used in [41] and [46] to address the problem of lack of training data. The basic idea is to use human input only where a classifier cannot provide a clear result, thereby significantly reducing the manual training process. A hybrid system is described in [25] which utilises both unsupervised (clustering) and supervised (instance-based learning and decision trees) machine learning techniques. High-dimensional overlapping clustering is applied in [34] as an alternative to traditional blocking (in order to reduce the number of record pair comparisons to be made), while in [27] the use of simple k-means clustering together with a user-tunable fuzzy region for the class of possible matches is investigated. Methods based on nearest neighbours are explored in [11], with the idea being to capture local structural properties instead of a single global distance approach. Graphical models [38] are another unsupervised technique that aims at using the structural information available in the data to build hierarchical probabilistic models for record pair classification.

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). Additionally, many of the recent publications in this area present experimental studies that are based on only small data sets with a couple of thousand records [12]. More work is required in this area to develop fully automated data linkage and geocoding techniques for very large data sets with millions of records.

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

- First, even when using blocking the computational requirements (memory usage and CPU 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.

- Second, comparing large number of record pairs also results in many pairs being classified as possible matches, and the manual clerical review process therefore becomes more time consuming, or even impossi-

ble. For the above example, if only 0.01% of the compared record pairs are classified as possible matches, manual review is required for $10,000$ 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.

- 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, for example in the health sector, where a breach of privacy could lead to a person's medical history being compromised. New application areas of data linkage, for example electronic health records stored on smart-cards that can be accessed by GPs and specialist doctors, public and private health insurers, as well as the national health administration system, or national security surveillance systems that link data from various government and private sources, 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. Privacy-preserving linking and geocoding techniques are required to allow the linking of large scale 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 approach 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 distributed privacy-preserving data linkage and geocoding of very large data collections possible. In the following section we start by illustrating the significance of privacy-preserving data linkage and geocoding by providing several real world scenarios, followed in Section 3 by a discussion of current techniques. The four research areas are then identified in Section 4, and we conclude this paper with a short discussion of further issues in Section 5.

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 various scenarios taken from real world situations.

Scenario 1: *An epidemiologist working at an university 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.* □

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 their database in order to conduct a spatial analysis of different types of cancer in their region. Due to limited resources the register cannot invest in an in-house geocoding system (i.e. software and personnel) but it 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. In order to obfuscate their data the register might

use *chaffing* [40] by adding dummy address records into their database. This will however increase the costs of geocoding, as commercial services charge according to the number of addresses they geocode. As an alternative, the register might be able to use the geocoding service of a trusted proxy organisation, like a government health department. In any case, 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 (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 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 of the third party with either company, or potential security breaches at the linkage party by intruders.

Scenario 4: *A national security agency is collecting and linking information from various data sources, including government and commercial databases, as well as public data as available on the Internet, with the aim to prevent crime, fraud and terrorism through surveillance of suspicious individuals. Even when done within a legal framework, there is limited public support of such a scheme due the possibility of potential misuse of the linked data.* □

In this scenario, techniques are needed that allow large scale linking of massive data collections with millions to billions of records between organisations without any of the data sources having to reveal any of their identifying data. Only in a situation where a set of records were matched (indicating suspicious behaviour) more detailed information can be released to the security agency. Equally important in this scenario is public understanding of how such privacy-preserving techniques work, and that they protect every individual's personal information while still allowing security agencies to track suspicious individuals in order to prevent crimes and terrorism.

Scenario 5: *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, possibly including data that is publicly available, that individually only contain de-identified data (i.e. data that does not allow re-identification).

Scenario 6: *A national census organisation aims to conduct longitudinal linkage of census data that is being collected at ten year intervals. However, national legislation dictates that names and addresses of all census records have to be destroyed within six months after the census date.* □

Traditional techniques would require that in this scenario only attributes other than names and addresses can be used for the linkage (like age, religion, profession, education, etc.), which will limit the quality of the linked data. Techniques that allow linking of encrypted and encoded data in such a way that approximate matching is possible, in combination with locking away the encryption key until the next census date (preferably to a trusted external organisation) might make longitudinal linkage in this scenario possible, both within the legal framework and the general public's acceptance.

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 privacy-preserving 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 [17, 31]. This, however, does little to obfuscate the source of those records. In many circumstances knowledge of the data source per-

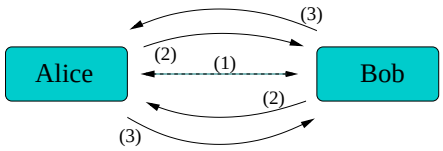

Figure 1. Basic two-party linkage protocol.

mits 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 [5], 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. 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 – an approach which has been termed *anonymous record linkage* [7]. 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. Truly anonymous data does not contain sufficient partially identifying information to permit any useful data linkage, by definition.

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 simple protocol based on this idea is described in [31], 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 [17], where medical details are separated from person identifiers and encrypted using different keys, and a highly trusted third party is responsible for obfuscating the sources of records before sending them to a single population register that is responsible for the linking of personal details and maintaining unique person identifiers.

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

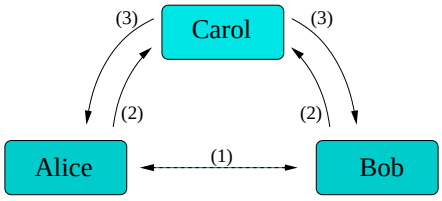

Figure 2. Basic three-party linkage protocol.

However, 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 (or anonymised) versions of the linked records can then be used for subsequent analysis. If the use of anonymised data is not feasible, then at worst only a small subset of records from each of the databases (the records that were matched) needs to be given to the researchers, in which case it may be feasible to obtain direct consent from the individuals concerned.

First methods based on cryptographic techniques that implement this idea were proposed in the mid-to-late 1990s by a team of French researchers [8, 24, 37]. These methods, which use keyed one-way hash encoding functions [43], allow the party undertaking the linkage to use all of 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 [42]. However, this work does not address the important issue of (typographical) errors which occur in most real world databases, and thus this approach is limited to exact matching only.

In general, cryptographic approaches to secure data linkage and data sharing can be classified into two- and three-party protocols, as illustrated in Figures 1 and 2.

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 [23] 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) [43] (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 [40] to their data in the form of dummy records). Each party then sends this encoded data to the other party.

(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 also 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 during the protocol 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, including TF-IDF (commonly used in information retrieval) and the Euclidean distance is discussed in [39]. This protocol is based on a stochastic scalar product, that is provably consistent and as secure as the underlying set-intersection cryptographic protocol it is using. Another two-party protocol for secure and private sequence comparisons based on the commonly used edit-distance approach is presented in [3]. It applies homomorphic encryption in such a way the 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 privacy-preserving 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 as illustrated in Figure 2.

(1) The two data sources 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, 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 back from Carol to the two 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 involved.

A protocol termed *blindfolded record linkage* based on q-grams is presented in [18]. It allows for approximate matching of values with typographical errors by calculating the Dice co-efficient [13] similarity measure between hash-encoded sets of q-grams. Weaknesses of this protocol include that Carol could mount a frequency analysis attack against the encrypted q-gram sets and compare them to frequencies in similar data (for example names taken from a telephone directory), while the 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 [18], including the last minute election of the linkage party from a collection of many functionally equivalent parties. Proof-of-concept code has shown the feasibility of this approach; however, the computational and communication overheads of encoded q-gram sets make the approach currently impractical for linking large real world data sets, or data containing long sequences such as those used in genomics or proteomics [18].

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 [36]. They are based on hash encoded values and improve the security weaknesses of [18]. 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 also have good security characteristics and minimise information leakage. They can, however, only perform exact matches and cannot deal with typographical errors and other variations in the data.

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. A first set of methods for privacy-preserving blocking has recently been presented in [2]. A secure three-party protocol based on hash encoded values (named *hash signatures*) and TF-IDF (similar to [39]) 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 (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.

3.4 Secure geocoding

Privacy-preserving data linkage techniques can also be useful for geocode matching. Traditionally, there are two basic approaches to geocoding. In the first, the data to be geocoded is sent by the data source to a (commercial) geocoding service, thereby compromising the privacy of all addresses in the data. In the second approach, the data source purchases the geocoding software and reference database, and then performs the geocode matching in-house. The advantage of this second approach is that no addresses have to be given to an external organisation; however, the disadvantages are the costs of purchasing the geocoding system and reference data, as well as training of in-house expertise in performing geocoding. Thus, the second approach will only be viable for large organisations, but prohibitive for research groups and non-government organisations like disease registers.

While geocoding is similar to data linkage [16], it is specific in that 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 an address is not available in the reference data, the location of the address should be extrapolated using the other addresses in the same street. Similarly, if an address cannot be found in its given postcode or suburb area, the linkage system should extended its search to neighbouring areas [16]. Most importantly, the locations of the linked addresses must not be revealed to the organisation undertaking the geocoding.

A privacy-preserving geocoding approach would allow an organisation to locally encrypt their address data and transfer them to a geocoding service, without having to reveal any of these addresses, and without the geocoding organisation learning which addresses are being matched. Several of the approaches presented in the previous subsections can be used for such a task; however, so far only in [18] 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 [19, 47] and *secure multi-party computation* [30, 32], as well as *secure information sharing* [1, 52], a field first introduced in 1982 [51]. Two-party protocols for minimal information sharing are presented in [1], where protocols for intersection, equijoin, intersection size and equijoin size across private databases are discussed and analysed. The challenges of privacy-preserving data sharing and integration are raised in [19], and a framework is presented in the contexts of databases and data mining. Distributed privacy-preserving data sharing in a system of autonomous entities is presented in [52]. The authors consider a *threat space* consisting not only of semi-honest but also malicious adversaries, and define measurements for information leakage, before proposing two-party protocols that can efficiently protect privacy. Configurable secure multi-party protocols are presented in [32]. They are based on quasi-communicative encryption using the concept of a one-way accumulator (a hash function that satisfies the quasi-communicative property). A semi-trusted centralised third party is assumed that performs the analysis of distributed data sent to it (but does not hold any data itself), allowing for secure centralised analysis of multi-party data in the face of malicious behaviour.

Although it appears that 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. [1, 32, 52] all consider the complexity of their protocols, but only [1] gives real world performance estimations (i.e. estimated run-times for example applications).

To summarise this overview, many of the presented approaches to 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. Cryptographic techniques often result in large computational and communication overheads, making the linkage of very large data set currently impossible. Additionally, none of the presented techniques has investigated the use of machine learning based automated record pair classification (as discussed in Section 1) within a privacy-preserving framework. In

the following section we identify and discuss four core research areas that need to be addressed to make the privacy-preserving linking and geocoding of very large data sets in distributed environments practical.

4. Research Directions

To the best of our knowledge, no work into the overall development of large-scale distributed privacy-preserving 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 [24, 31] or restrict the way linkage can be performed [7], while other methods only allow exact matching [36, 37, 42].

Only in the last three years have methods been developed that allow approximate matching without the need of the original values having to be revealed to other parties [2, 3, 18, 39]. 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 [26, 49], nor using one of the recently developed machine learning based techniques (as discussed below).

Research in this area should aim to develop frameworks that allow the inclusion of a wide variety of secure approximate string comparisons techniques [13], including the commonly used Jaro and Winkler comparators [49], which so far have not been converted into a privacy-preserving setting [18]. 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 privacy-preserving linkage that have reduced communication and computational overheads compared to current methods, as otherwise linking large data sets will be problematic. Secure approaches for both two- and three party protocols are needed for a large number of similarity comparison techniques [13] 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 privacy-preserving blocking [2] so that linking of very large data sets will become feasible. Modifying the developed protocols and methods so that privacy-preserving geocode matching can be performed will also be of importance.

As there is often a trade-off between privacy and performance (increased privacy preservation normally comes at higher encoding and communication costs), it is crucial to develop theoretical frameworks that allow better understanding of this trade-off, as well as methods that offer different levels of privacy preservation to allow the use of linking in applications that require different levels of security.

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 methods based on machine learning, artificial intelligence and information retrieval techniques have been developed in the past few years [6, 10, 11, 20, 21, 22, 25, 27, 34, 35, 38, 41, 46, 50, 53]. However, none of these methods takes 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 [22, 25, 27] and hierarchical graphical models [38] have shown to be promising in the context of data linkage, no work has so far been done to use such techniques within a secure setting. Unsupervised techniques have to be reconsidered from a privacy preservation point of view. For clustering algorithms, for example, the question of how to calculate distances using only encoded values has to be solved. Assessing the quality of privacy-preserving linkage techniques is another challenge. Techniques developed in privacy-preserving data mining [47] and machine learning 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.

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 (for example a researcher). 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).

Only limited research has so far been done in this area [14, 28, 42]. Some recent work has shown that parallel data linkage can achieve good speedup results [14], 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 parallel and distributed platforms like clusters or grids, or when linkage is done between different 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 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 privacy-preserving 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.

Another important issue that is related to scalability is the availability of large test data collections that will allow testing, evaluation and comparisons of new algorithms and techniques. This has so far only been addressed to a limited extent. To the best of our knowledge, the *RIDDLE*[2] (Repository of Information on Duplicate Detection, Record Linkage, and Identity Uncertainty) is the only initiative towards this so far. As privacy and confidentiality issues make it unlikely that real data containing names and addresses will ever be made publicly available, synthetic data has to be used. Developing data generators that are able to create realistic personal information is a challenging task itself [12, 29]; however, using synthetic data has the advantages that its content and error characteristics can be controlled, that the deduplication or linkage status is known, and that such data can easily be made publicly available to other researchers.

[2] URL: `http://www.cs.utexas.edu/users/ml/riddle/`

4.4 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 be made useless. As shown in Scenario 5 in Section 2, while properly de-identified linked data 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 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 [48], techniques for masking data (like swapping or aggregating values) so that the data can be made public while guaranteeing no re-identification will be possible. In the security and data mining communities initial work on k-anonymity [44] and trail re-identification [33] will have to be further investigated, with the aim to fully integrate such techniques 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. We then 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.

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 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 data linkage and geocoding techniques, especially privacy-preserving approaches such as the ones presented in this paper, 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 privacy-preserving 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. The author would like to thank Paul Thomas for helpful comments and proofreading.

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] R. Baxter, P. Christen, and T. Churches. A comparison of fast blocking methods for record linkage. In *Proceedings of 9th ACM SIGKDD Workshop on Data Cleaning, Record Linkage and Object Consolidation*, Washington DC, 2003.

[5] J. J. Berman. Confidentiality issues for medical data miners. *Artificial Intelligence in Medicine*, 26:25–36, 2003.

[6] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of ACM SIGKDD*, pages 39–48, Washington DC, 2003.

[7] T. Blakely, A. Woodward, and C. Salmond. Anonymous linkage of New Zealand mortality and census data. *ANZ Journal of Public Health*, 24(1):92–5, 2000.

[8] H. Bouzelat, C. Quantin, and L. Dusserre. Extraction and anonymity protocol of medical file. In *AMIA Fall Symposium*, pages 323–327, 1996.

[9] E. Brook, D. Rosman, D. Holman, and B. Trutwein. Summary report: Research outputs project, WA Data Linkage Unit (1995-2003), 2005. Western Australia Data Linkage Unit, Perth.

[10] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani. Robust and efficient fuzzy match for online data cleaning. In *Proceedings of ACM SIGMOD*, pages 313–324, San Diego, 2003.

[11] S. Chaudhuri, V. Ganti, and R. Motwani. Robust identification of fuzzy duplicates. In *Proceedings of the 21st international conference on data engineering (ICDE'05)*, pages 865–876, Tokyo, 2005.

[12] P. Christen. Probabilistic data generation for deduplication and data linkage. In *IDEAL'05*, pages 109–116, Brisbane, 2005. Springer LNCS 3578.

[13] P. Christen. A comparison of personal name matching: Techniques and practical issues. Submitted to IEEE International Conference on Data Mining (ICDM), 2006.

[14] 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.

[15] 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.

[16] 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.

[17] 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. Available online at: http://www.biomedcentral.com/1472-2288/3/1/.

[18] T. Churches and P. Christen. Some methods for blindfolded record linkage. *BioMed Central Medical Informatics and Decision Making*, 4(9), 2002. Available online at: http://www.biomedcentral.com/1472-6947/4/9/.

[19] C. Clifton, M. Kantarcioglu, A. Doan, G. Schadow, J. Vaidya, A. Elmagarmid, and D. Suciu. Privacy-preserving data integration and sharing. In *DMKD '04: Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 19–26. ACM Press, 2004.

[20] W. W. Cohen. Integration of heterogeneous databases without common domains using queries based on textual similarity. In *Proceedings of ACM SIGMOD*, pages 201–212, Seattle, 1998.

[21] W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proceedings of IJCAI-03 workshop on information integration on the Web*, pages 73–78, Acapulco, 2003.

[22] W. W. Cohen and J. Richman. Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of ACM SIGKDD*, pages 475–480, Edmonton, Canada, 2002. ACM Press.

[23] W. Diffie and M. E. Hellman. New directions in cryptography. *IEEE Transactions on Information Theory*, IT-22(6):644–654, 1976.

[24] 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.

[25] M. G. Elfeky, A. K. Elmagarmid, and V. S. Verykios. TAILOR: A record linkage tool box. In *ICDE '02: Proceedings of the 18th International Conference on Data Engineering*, pages 17–28, San Jose, 2002. IEEE Computer Society.

[26] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Society*, 64(328):1183–1210, 1969.

[27] L. Gu and R. Baxter. Decision models for record linkage. In *AusDM'04, Springer LNAI 3755*, pages 146–160, Cairns, Australia, 2006.

[28] 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.

[29] M. A. Hernandez and S. J. Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, 2(1):9–37, 1998.

[30] M. Hirt, U. Maurer, and B. Przydatek. Effi cient secure multi-party computation. In *Proceedings of ASIACRYPT, LNCS 1976*, pages 143–161, 2000.

[31] 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.

[32] B. Malin, E. Airoldi, S. Edoho-Eket, and Y. Li. Confi gurable security protocols for multi-party data analysis with malicious participants. In *ICDE '05: Proceedings of the 21st International Conference on Data Engineering (ICDE'05)*, pages 533–544, Tokyo, 2005. IEEE Computer Society.

[33] 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.

[34] A. McCallum, K. Nigam, and L. H. Ungar. Effi cient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of ACM SIGKDD*, pages 169–178, Boston, 2000.

[35] U. Y. Nahm, M. Bilenko, and R. J. Mooney. Two approaches to handling noisy variation in text mining. In *Proceedings of the ICML-2002 workshop on text learning (TextML'2002)*, pages 18–27, Sydney, 2002.

[36] 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.

[37] C. Quantin, H. Bouzelat, F. Allaert, A. Benhamiche, J. Faivre, and L. Dusserre. How to ensure data quality of an epidemiological follow-up: Quality assessment of an anonymous record linkage procedure. *International Journal of Medical Informatics*, 49(1):117–122, 1998.

[38] P. Ravikumar and W. W. Cohen. A hierarchical graphical model for record linkage. In *roc. of the 20th Conference on Uncertainty in Artifi cial Intelligence*, pages 454–461, Banff, Canada, 2004.

[39] 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.

[40] R. L. Rivest. Chaffi ng and winnowing: Confi dentiality without encryption. MIT Lab for Computer Science. Available at: http://theory.lcs.mit.edu/∼rivest/chaffi ng.txt.

[41] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of ACM SIGKDD*, pages 269–278, Edmonton, 2002.

[42] 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.

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

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

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

[46] S. Tejada, C. A. Knoblock, and S. Minton. Learning domain-independent string transformation weights for high accuracy object identifi cation. In *Proceedings of ACM SIGKDD*, pages 350–359, Edmonton, 2002.

[47] 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.

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

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

[50] W. E. Yancey. An adaptive string comparator for record linkage. Technical Report RRS2004/02, US Bureau of the Census, 2004.

[51] A. C. Yao. Protocols for secure computations. In *Proc of 23rd IEEE Symposium on the Foundations of Computer Science (FOCS)*, pages 160–164, 1982.

[52] N. Zhang and W. Zhao. Distributed privacy preserving information sharing. In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pages 889–900. VLDB Endowment, 2005.

[53] J. J. Zhu and L. H. Ungar. String edit analysis for merging databases. In *KDD workshop on text mining, held at ACM SIGKDD*, Boston, 2000.