Geocode Matching and Privacy Preservation

Peter Christen

Department of Computer Science, The Australian National University
Canberra ACT 0200, Australia
`peter.christen@anu.edu.au`

Abstract. Geocoding is the process of matching addresses to geographic locations, such as latitudes and longitudes, or local census areas. In many applications, addresses are the key to geo-spatial data analysis and mining. Privacy and confidentiality are of paramount importance when data from, for example, cancer registries or crime databases is geocoded. Various approaches to privacy-preserving data matching, also called record linkage or entity resolution, have been developed in recent times. However, most of these approaches have not considered the specific privacy issues involved in geocode matching. This paper provides a brief introduction to privacy-preserving data and geocode matching, and using several real-world scenarios the issues involved in privacy and confidentiality for data and geocode matching are illustrated. The challenges of making privacy-preserving matching practical for real-world applications are highlighted, and potential directions for future research are discussed.

Keywords: Data matching, record linkage, entity resolution, privacy preservation, geocoding, secure multi-party computations.

1 Introduction

Increasingly large amounts of data are being collected by many business organisations, government agencies, and research institutes. A large portion of this data is about people, for example patients, students, clients and customers, travellers, or tax payers. Commonly, personal details (such as names, addresses, dates of birth, telephone, driver's license and social security numbers) are stored in databases together with application specific information, for example medical details, student enrolments and grades, customer orders and payments, travel and immigration details, or tax payments. Most of this information is considered to be private or confidential, and appropriate laws and regulations are in place in many countries that assure such information is protected properly, and that data holders do not publish such information or disclose it to others.

In our networked world, however, where many of us gather information and interact with each other and various organisations online, there is both a need and a desire for services that allow publication of information by both individuals and organisations. Examples include social networking and blogging Web sites where individuals can publish ideas, comments, profiles, photos and videos; or online mapping services that allow users to search for locations and addresses,

that enable the uploading of photos taken at certain locations, or that show the geographic locations of recent news events. While many of these online services are useful, informative and entertaining, they can also pose a threat to privacy, because they facilitate the way in which private or confidential pieces of information can be accessed, matched and made public.

It was recently estimated that around 80% to 90% of all governmental data collections contain details of some kind of geographic locations [1]. In many cases, these locations are addresses, and they are the key to spatially enabled data, for example to match personal information to geographic locations. The aim of such *geocode matching* is to generate geographic locations (such as latitudes and longitudes) from street address information, as will be discussed in more details in Sect. 2.1. Once data has been geocoded, it can be used for spatial data mining, and it can be visualised and combined with other data using Geographical Information Systems (GIS) and online mapping services.

In the health sector, geocoded databases can be used to detect local clusters of diseases, and many environmental studies rely upon geocoded data and GIS to map and visualise areas of possible exposure to health risks, and to locate people who live in relation to these areas. Geocoded data can help businesses to better plan marketing and future expansion, for example where to locate new stores or supply centres. In national censuses, geocoded data can be used to assign households to local census areas, which are often the basis of a variety of statistical data analysis projects.

The remainder of this paper is structured as follows. In the next section, a general overview of *data matching* is given, while *geocode matching* is discussed in Sect. 2.1. The technique of *reverse geocoding* is then described in Sect. 2.2, and the issues of privacy and confidentiality of data matching in general are the topic of Sect. 2.3. In order to illustrate these issues, several data and geocode matching scenarios are given in Sect. 3. Various privacy-preserving data matching techniques have been developed in the past few years, and an overview of these techniques is provided in Sect. 4. Finally, the paper is concluded in Sect. 5 with an outlook to potential research directions.

2 Data Matching

Data matching is the process of linking and aggregating records that refer to the same entity from one or more databases [2, 3]. A variety of techniques for data matching have been developed in different fields in the past, and while computer scientists speak of *data* or *record matching*, or *entity resolution*, statisticians and health researchers refer to *data* or *record linkage*, and the database and business oriented IT communities call this process *data cleaning* or *cleansing*, *ETL* (extraction, transformation and loading), *object identification*, or *merge/purge processing*. When the aim is to find duplicate records (i.e., records that refer to the same entity) in only one database, then this process is often called *duplicate detection*, *deduplication*, or *internal data linkage* [4].

Traditionally, data matching has been used in the health sector for matching epidemiological and administrative databases for research purposes [5], and by census agencies to create data sets that allow the production of a variety of statistical analyses [3]. Today, data matching techniques are applied in an increasing number of both private and public sector organisations. Many businesses routinely deduplicate their customer databases, and match them with data obtained from other sources, for example for marketing purposes. Government agencies match databases to detect fraud and improve outcomes in taxation, immigration and social welfare, while security agencies conduct matching of databases from a variety of sources with the objective to assemble crime and terrorism intelligence [6, 7]. Another application that increasingly relies upon data matching is the assembly of bibliometric impact data, where the aim is to collect all publications of researchers and their corresponding citation counts [8].

Commonly, no unique entity identifiers are available in all the databases to be matched. Therefore, the attributes that identify entities need to be used for the matching. In many cases, these attributes are personal identifiers, such as names, addresses, dates of birth, and telephone or social security numbers. The matching process is usually challenged because real world data is dirty [9]: values commonly contain typographical errors and variations, they can be out of date or missing, and the databases to be matched might even use different coding schemes. Cleaning and standardising databases before matching is attempted is therefore a crucial pre-processing step [10], with the main tasks being the conversion of the raw input data into well defined, consistent forms, and the resolution of representation and encoding inconsistencies.

The process of matching consists of three major steps: blocking or indexing, record pair comparison, and classification of the compared record pairs.

1. Blocking or indexing

When two databases are matched, potentially each record in the first database should be compared with all records in the second database. Thus, the total number of potential comparisons is of quadratic complexity. On the other hand, most of these comparisons correspond to non-matches, because the maximum number of matches can only be in the order of the number of records in the smaller of the two databases to be matched (assuming these databases do not contain duplicate records) [2]. So, while the computational efforts potentially increase quadratically with the size of the databases to be matched, the maximum number of true matches only increases linearly.

The main performance bottleneck in the data matching process is usually the expensive detailed comparison of attribute values between records (as discussed below) [2, 11], and therefore it is not feasible to compare all record pairs when the databases are large.

Techniques commonly known as *blocking* [11] are therefore applied, with the aim to reduce the actual number of record pairs that are to be compared. These techniques work by indexing or clustering the databases into blocks, such that candidate pairs of records, which are likely to correspond to a true match, are inserted into the same block. Pairs of records are then

generated only from the records that are in the same block. While blocking will remove many of the obvious non-matching record pairs, some true matches will likely also be removed in the blocking step, because of errors or variations in record attribute values [2].

2. Record pair comparisons

Selected attributes (or fields) of the candidate record pairs generated in the blocking step are compared using a variety of comparison functions [12, 13]. To account for typographical errors and variations, strings are normally compared using approximate string comparison functions [14], while specific comparison functions exist for other data types, such as ages, times, dates, or telephone, social security or credit card numbers [12].

All these similarity functions return a numerical *matching weight*, which is commonly normalised, such that an exact match between two attribute values returns 1.0, while the comparison between two completely different values returns 0.0. For each compared record pair, a *weight vector* is generated, which contains one matching weight for each compared attribute. For example, if surnames, given names and postcodes are compared, then three matching weights would be calculated for each compared record pair.

3. Record pair classification

In the traditional data matching approach [3], for each compared record pair, the weights in its corresponding weight vector are summed into one total matching weight. Using two thresholds, a record pair is then classified as a *match*, a *non-match*, or as a *possible match* (these are the record pairs for which manual *clerical review*, a time-consuming and cumbersome process, is required to decide their final match status). In the past decade, researchers from the fields of data mining, machine learning, artificial intelligence and databases have explored the use of various techniques with the objective to make the classification of record pairs both more accurate and more automatic [2, 3]. Many of these novel techniques are based on supervised machine learning approaches and therefore require training data, which unfortunately is often not available in real world applications [15].

Matching today’s increasingly large databases has several challenges. First, as already mentioned, real-world data is dirty, and matching therefore is reliant upon the attributes available in common in all databases to be matched. In real-world applications, data quality can be one of the biggest obstacles to successful data matching [16]. Second, even when efficient blocking techniques are applied, the number of candidate record pairs can be very large, and the resulting computational requirements (run-times and memory foot-prints) can become challenging even for today’s powerful computing platforms.

Finally, because the matching is often based on personal information, like name and address details, privacy and confidentiality become of paramount concern, especially when databases are matched between different organisations. This challenge will be discussed in more detail in Sect. 2.3.

Fig. 1. Example geocoding using property centres (numbers 1 to 9) and street centre-lines (the thick dark/red lines and numbers 10 to 17, with the dotted lines corresponding to a global street offset).

2.1 Geocode Matching

Geocode matching, or *geocoding*, is the process of matching user address data with a reference database that contains cleaned and standardised addresses and their geographic locations [17]. Geocoding is a special case of data matching, and it is challenged by the same issues as data matching, as discussed above. Specifically, addresses are often dirty, in that they can include misspellings of street and suburb names, can have wrong or incomplete street or apartment numbers, wrong zipcodes, or parts of an address can even be missing. It has been reported that a geocode matching rate of around 70% is acceptable when user addresses have been recorded over the telephone, have been scanned, or were manually typed from hand-written forms [18].

Additionally, addresses can quickly become out-of-date as people move to new homes. This is especially an issue with younger people who are much more mobile than older generations. As new suburbs are being built, and postal agencies adjust their zipcode areas, new addresses are becoming valid, while parts of existing addresses are changing and some addresses even become invalid. Matching a user address to the correct reference address can therefore become quite challenging. Scalability and computational efforts are also of concern, as a reference address database usually contains many million addresses, and so often does a user's database. Efficient blocking techniques are thus important.

As illustrated in Fig. 1, there are two basic geocoding techniques, which depend upon the available reference database. Street level geocoding is based on a reference database that contains street centreline data, which are the geographic locations of the start- and end-points of street segments, as well as the range of street numbers they cover. A popular, freely available such database for the USA is the *Tiger* (Topologically Integrated Geographic Encoding and Referenc-

Fig. 2. Reverse-geocoding example. The dots on the map could, for example, correspond to certain disease cases or the locations of certain crime types.

ing) system by the US Census Bureau.¹ With this type of geocoding, positional errors are introduced because street segments normally do not follow the exact location of streets, the locations of street numbers are calculated assuming regular intervals between properties, and commonly a global street offset (distance from the street centreline) is used to calculate the actual geographic locations.

An alternative geocoding approach is possible if a property reference database is available that contains the centre of properties, or even buildings, for all addresses in an area. As several recent studies have shown [17, 19], this approach can result in more accurate geocoding, especially in rural areas where properties and street numbers are not evenly spaced, and where street centreline data is less accurate than in urban areas. Even small differences in the geocoded locations can for example result in addresses being assigned to different local census areas, and this can have significant implications for any following spatial analysis that is based on such local census areas [17].

2.2 Reverse Geocoding

Reverse geocoding is the process of matching a given geographic location (usually provided as latitude and longitude) back to an address (or another entity, such as a building or property). The geographic location can be extracted from maps, such as the example shown in Fig. 2, which can often be taken from Web sites or from publications. One challenge with reverse geocoding lies in accurately positioning the map relative to its true geographic location to enable accurate matching of locations to addresses. Another issue is related to the resolution of a given map, i.e., the physical area covered by a single pixel, and the way locations are marked on a map (for example as single highlighted pixels, as coloured circles, or as much larger illustrative markers).

¹ Available from: <http://www.census.gov/geo/www/tiger/>

Two recent studies [20, 21] illustrated the accuracy of reverse geocoding. They were based on maps with a resolution commonly used in publications, such as medical journals. With a higher resolution image file (226 dots per inch, and a scale of 1:100,000), the authors were able to accurately identify 432 of 550 (79%) randomly selected addresses in Boston within 14 meters of the real address [21]. With a lower resolution image file (50 dots per inch and scale 1:100,000), 144 of 550 (26%) addresses could be identified directly, while on average the real addresses were only 21 meters away from the reverse geocoded addresses. Overall, 99.8% of all addresses were within 70 meters of their real locations [20].

A similar analysis was conducted after hurricane Katrina using a map published by a local newspaper [22]. This map showed the locations of mortalities, and applying reverse geocoding, the authors were able to identify the actual locations of many of these mortalities. Using field teams with GPS receivers that recorded actual locations allowed the authors to measure the accuracy of their reverse geocoding process. They found that up to 40% of the reverse geocoded mortalities were located within 20 meters of the field verified actual residences.

2.3 Privacy and Confidentiality Issues

Data and geocode matching are technologies that can be employed in a variety of applications. For example, the aim of matching health databases is normally to investigate the effects of various aspects of the health system and how they interact with environmental or social factors, with the objectives to increase our understanding of these complex interactions, to improve the public health system, and to reduce costs to both patients and governments. Thus, this kind of data matching is aimed at increasing the general public's wellbeing. Similarly, when data matching is used by statistical agencies as part of the census, or by other government agencies to detect fraudulent behaviour, such as people registering for unemployment money while they are actually employed, then the public is generally supportive of these types of data matching projects.

On the other hand, when data matching is used by national security agencies, for example to assemble terrorism watch lists, then the public increasingly worries about the resulting privacy and confidentiality implications, because individuals can directly be affected [6]. A false match can result in an innocent individual being added to a list of suspected terrorists or criminals, resulting in serious implications.

To a similar degree, the public is also worried about how private organisations are selling, exchanging and matching their customer data, in order to improve their business or target specific groups for marketing. In many countries, consumers have only limited control over what private organisations can do with the data they have collected about them during their business activities.

Data matching techniques can also be used by criminals, for example for matching disparate pieces of information they collect as part of their efforts to conduct identity fraud. With an increasing amount of data about various aspects of people's life being available online, matching such disparate data with the aim to build somebody's identity becomes increasingly attractive, as this can be done

anywhere in the world with relatively little efforts, and does not require physical access to a victim's information (like searching through somebody's garbage with the aim to find identifying details) [23]. Such *re-identification* of individuals from using only publicly available, de-identified (or partially identifying) data has recently attracted interests from several research fields, and a variety of anonymisation techniques for data publishing have been proposed [24–26].

Geocode matching, and especially reverse geocoding, can have serious privacy implications as well. As the studies described in Sect. 2.2 showed, it can allow matching of information about people to the location where they live, potentially indicating an individual's vulnerability, such as somebody living alone or having a certain illness. Even when areas and not individual addresses are associated with certain characteristics, for example higher prevalence of a certain disease or higher crime rates, then this will affect all people living in such an area, and can result in residents and businesses avoiding or leaving the area. Health research and crime analysis are two domains that increasingly publish maps and images with geocoded data in research publications or on Web sites.

It is therefore paramount that the personal details of individuals are protected when data is matched or geocoded [27, 28]. This is especially important in applications where a breach of privacy can seriously impact an individual's financial, employment, insurance and social status. Data matching and geocoding will only receive widespread acceptance if privacy and confidentiality of the individuals whose data is being matched can be guaranteed. A variety of such approaches have recently been developed, as will be discussed in Sect. 4.

3 Data and Geocode Matching Scenarios

In this section, the previously discussed privacy and confidentiality issues are illustrated using several scenarios. The first two have been adapted from [27, 29], while the third scenario is based on ideas given in [30].

Scenario 1: Data Matching

A public health researcher is planning to investigate the effects that serious car accidents have upon the public health system. She is interested in what the most common types of injuries are, how much they cost the public health system, and how the general health of car accident victims is progressing after their recovery. For her analysis, the researcher requires access to data from public and private hospitals, general practitioners and specialist doctors, private health insurance companies, the police, and car insurance companies.

In most countries it is very unlikely that all these databases have a common entity identifier that enables exact matching of all records that refer to the same individual. Therefore, for this project, personal details are required for the matching. If the researcher is successful in getting access to all the required databases, then the matching can be performed by the researcher (or a support entity at the researcher's organisation) following strict security and access procedures. Alternatively, all databases could be given to a trusted organisation,

such as a data matching unit within a government health department, which would perform the matching and only provide the matched records without any identifying information (name and address details) to the researcher.

In both these cases, however, the original records (including the identifying personal details) from all databases required for this project have to be made available to the organisation that performs the matching. This requirement will very likely prevent some of the organisations involved from providing their data towards this project, and therefore prevent a research project that is of significant benefit to the public.

Scenario 2: Geocode Matching ²

In many countries, cancer registries collect data about all occurrences of cancer in a certain state or territory. Often, these registries are small organisations that are partially supported by government funding and public donations.

In this scenario, it is assumed a cancer registry would like to geocode its database to conduct spatial analyses of its cancer data. This would, for example, allow the registry to detect if there are clusters of cancers in certain areas, or correlations between increased cancer occurrences and environmental or socio-economic factors. Due to its limited financial resources, the cancer registry cannot afford to invest in an in-house geocoding system (consisting of software, as well as trained personnel), but instead is required to employ an external geocoding service provider.

The regulatory or legal framework under which the cancer registry operates might not allow it to provide its detailed data to any external organisation for geocoding. Even if this would be allowed, complete trust is required in the capabilities of the external geocoding service provider to accurately match the registry's data, and to then properly destroy all data received by the cancer registry or generated during the geocoding project. Additionally, if the geocoding service provider is a commercial organisation, then normally only limited independent information is available about its matching quality. Ideally, comparative studies, conducted by an independent organisation like a government agency, about the matching quality of various commercial geocoding products and service providers should be available to help the cancer registry select a suitable geocoding provider.

In order to obfuscate their data, the cancer registry might use *chaffing* [32], by adding dummy address records into its cancer database before sending it to the geocoding service provider. This will however increase the costs of geocoding, as a commercial provider will likely charge according to the number of addresses to be geocoded. As an alternative, the cancer registry might use the geocoding service of a trusted proxy organisation, such as a government health department. However, in both approaches, the original address details of the cancer patients have to be made available to the outside organisation that performs the geocoding, and this might pose a serious privacy risk.

² See [31] for a more general discussion of geocoding in cancer research.

Scenario 3: Reverse Geocoding

A police department is publishing crime statistics of its local area on its newly designed Web site. This Web site allows the public to select a variety of crime types (like burglaries, traffic offences, assaults, etc.) and time periods when they happened. It also allows some limited drill-down into smaller areas, such as suburbs. These crime maps initially improve the relationship between the police and the public, because they raise awareness where problematic neighbourhoods are, and because they allow the public to know where crimes and security events occur. As a result, the public gains confidence in the work the police is doing, and people are also more willing to work with the police.

On the other hand, these published crime maps are also misused. For example, vendors of burglary alarms target local areas with higher incident rates of burglaries, while the property values in these areas have dropped significantly since the crime maps have been published, because it is harder to attract new residents into these areas. Some residents even leave these areas because of the published crime statistics, and as a result, local businesses also consider to move to different, safer locations.

The published crimes maps also allow criminals to identify areas that have not received a lot of police attention in recent times, for example because the crime incident rates in these areas are low. Once criminals obtain socio-economic data for these areas from a government census agency, they find that some of these areas are suburbs with high average income and a large number of single residents. This indicates to these criminals that there is an increased number of potentially easy and lucrative targets in these areas.

Besides summarised crime statistics, the police Web site also allows the generation of ‘pin’ maps (as illustrated in Fig. 2), which show the exact locations of crimes in recent times. Filtering and selection options allow visualisation of serious crimes, such as assaults, murders and sexual offences. This results in several victims being contacted by media organisations, which re-traumatises them. Following from these incidents, the police department sees a significant reduction in the number of reports in sexual offences, as victims decide not to report an assault because of fear of publicity.

4 Privacy-Preserving Matching

Traditionally, when databases from different organisations are to be matched, then the organisation that undertakes the matching will require access to all the records from all the databases to be matched, because it is impossible to know before the matching which records will match. Complete trust is required in the intentions of all organisations involved in a matching project, in their ability to maintain confidentiality, in the security of their networking and computing facilities, and in the reliability of the staff involved in matching projects.

In situations where the data is matched for a research project, for example, the matching is often conducted by a trusted organisation, such as a government regulated matching unit. For the matching of health related databases, good

Fig. 3. Basic two- and three-party protocols for privacy-preserving data matching. Alice and Bob are the data holders, and Carol is the trusted data matching unit.

practise dictates that all medical details are removed from the data before it is given to such a matching unit [5, 33], and that the researcher is only given the required medical details of the matched records, but not the full identifying attributes. With this approach, however, even only knowing the source of the data to be matched might permit highly confidential information to be inferred about individuals who are identified in the databases to be matched.

However, this invasion of privacy could be reduced or even avoided if there were some approach that allows the detection of records that refer to the same entity in databases from different organisations, without either organisation having to reveal any identifying details to any other organisation. Such approaches have been developed in the past few years, most of them based on cryptographic and secure multi-party computation techniques.

Pioneering work was done in the 1990s by French health researchers [34–38]. Their approach was based on keyed one-way hash encoding functions [39], which allowed matching using only encoded identifying data. These approaches provided good privacy protection against a single organisation trying to find out information about the other organisation’s data. However, a major problem when using hash-encoded values (for example, the original un-encoded value ‘tim’ could be encoded into the hash-string ‘51d3a6a70’) is that a single character difference in an original value results in a completely different hash encoding, and thus only exact matches will be found. Applying phonetic encoding, such as Soundex or NYSIIS [14], to the original values before they are hash-encoded can help to overcome some variations, but only to a limited degree.

Cryptography based privacy-preserving data matching approaches can generally be classified into two- and three-party protocols, as illustrated in Fig. 3. Within a two-party protocol, the two data holders, *Alice* and *Bob*, plan to match their databases such that only information about the matched records is revealed to each other. The following principle three steps are involved: (1) The data holders agree upon a secret random key which will be used to encode their data in the following steps. It is also assumed that all communication is conducted in an authenticated and secure fashion, for example using a public key infrastructure (PKI) [39]. (2) The two data holders pre-process and encode their databases in an agreed fashion, and then send their encoded database to each other. (3) Each data holder performs the matching using their own data and the encoded data they received from the other data holder, and thus both data holders learn

which records match. The two data holders can then exchange their matching results and negotiate how to proceed next. Steps (2) and (3) might be repeated several times, depending upon the actual matching technique employed. The most important requirement for any two-party data matching protocol is that at any time neither data holder will have all the information required to infer the values of the original data of the other data holder.

One recently proposed two-party protocol for string distances is based on a stochastic scalar product, that is as secure as the underlying set-intersection cryptographic protocol it uses [40]. Another approach is aimed at secure sequence comparisons [41], and is based on edit-distance calculations. It applies encoding such that neither party at any time has the complete information about the dynamic-programming matrix used for the edit-distance calculation.

Three-party protocols for privacy-preserving data matching are based on the idea that a trusted third party, commonly called *Carol*, conducts the matching in such a way that neither of the two data holders have to provide any private or confidential information to any other party involved in the protocol, and *Carol* only sees encoded values. The general three-party protocol also consists of three principal steps, as illustrated in Fig. 3: (1) The two data holders again agree upon a secret random key which will be used to encode their data in the following steps. Note that this key is not shared with *Carol*. (2) The two data holders now pre-process and encode their databases in an agreed fashion, and then send their encoded database to *Carol*, which performs the matching without seeing any of the original values. That is, *Carol* must perform the matching on the encoded data she received from *Alice* and *Bob*. (3) Once the matching is completed, *Carol* sends information about the matched records back to both data holders. Depending upon what was agreed, this might only be the number of matched records, or it might be their record identifiers. Depending upon the outcome of the matching, *Alice* and *Bob* then negotiate how to proceed next.

Several three-party protocols have recently been developed. They mainly differ in the way the matching party is calculating the similarity between encoded record attribute values, and by how much information can be inferred by any of the parties involved in the matching protocol.

Two protocols that not only enable data matching, but also allow cohort extraction, have recently been proposed [42]. Combined, they facilitate the construction of a matched data set in such a way that no identifying information is revealed to any other party involved, and neither of the data holders learns which of their records have been extracted from their databases. These two protocols are based on hash-encoded values. They can only perform exact matching, and are thus of limited use when the data contains errors and variations.

Another three-party protocol, named *blindfolded record linkage* [43], is based on hash-encoded q -grams (sub-strings of length q), and allows approximate matching by calculating the *Dice* co-efficient similarity measure between hash-encoded sets of q -grams. The major drawback of this approach is its computational and communication overhead, which makes it impractical for large databases or long strings, such as suburb names or genome sequences.

As discussed in Sect. 2, one major challenge when matching large databases is the potential number of record pairs that need to be compared. Blocking techniques are required to make large-scale matching possible [11], however, none of the so far presented privacy-preserving matching approaches take blocking into account. A recently presented approach to secure blocking [44] is based on a three-party protocol, and uses hash-encoded values and secure string distance calculations, similar as used in other approaches [40]. The basic idea of the approach is to compare records between databases only if they have at least one token in common (for example a word or q -gram).

A different approach that was recently proposed [45] is aimed at improving the performance of privacy-preserving data matching by using a hybrid technique which combines sanitisation techniques (like k -anonymity [26]) with cryptographic secure multi-party computations (SMC). The idea of this approach is to first use the anonymised data sets to decide the match status (match or non-match) of a large proportion of record pairs, and secondly to only use SMC for the remaining, hard to match pairs. The first step of this approach can be seen as a blocking step.

4.1 Privacy-Preserving Geocoding

To the best of the author’s knowledge, only one publication has so far considered privacy-preserving geocode matching [43]. What is required is that, besides the data holder, no party involved in a privacy-preserving geocode matching protocol should be able to learn which addresses were matched, as otherwise the geographic locations of the data holder’s addresses would be revealed. Thus, a method for privacy-preserving geocode matching should allow a data holder to locally encode their address records and then transfer them to a geocoding service provider, without having to reveal any of their addresses, and without the geocoding service provider learning anything about these addresses.

This process is similar to the cohort extraction protocol discussed above [42], however, because addresses are often dirty (i.e., contain errors and variations), approximate matching techniques are required. A variation of the q -gram based three-party protocol [43] does allow such a privacy-preserving geocoding, but as discussed before, the computational and communication overheads of this approach prohibits the geocoding of large databases.

Besides variations and errors in user address data, geocoding, when based on a property reference address database, also has to deal with issues such as user street addresses that are not available in the reference database. In such cases, the location of the missing street number needs to be extrapolated using the closest street numbers that exist in the reference database. Similarly, if a given user address cannot be found in its expected zipcode or suburb area, the matching should be extended to neighbouring areas, because people commonly provide neighbouring zipcodes, especially if they have a higher social status [17]. Many of the approaches to privacy-preserving data matching presented in Sect. 4 can be used as starting points to develop privacy-preserving geocode matching.

4.2 Challenges

Many of the presented approaches to privacy-preserving data matching are in a proof-of-concept or prototype state, and they currently only allow matching of small to medium sized databases, or they can only perform exact matching. Before privacy-preserving data and geocode matching can be employed in practice, the following challenges need to be addressed.

- **Improved secure matching techniques**

Approaches that enable approximate matching of databases in a privacy-preserving way have only been developed in the past few years [40, 41, 43, 44]. Using secure multi-party computations, they securely compute functions at the expense of computational and communication overheads. Many of these approaches are of quadratic or even higher complexity, and are therefore of limited scalability. Additionally, while they allow approximate matching, they are only partial solutions, in that they are not integrated into either the traditional probabilistic data matching approach, nor into one of the various recently developed machine learning based techniques [3, 29].

- **Automated record pair classification and quality assessment**

A variety of novel data matching classification methods has been developed in the past few years [2, 3, 15], however, none of them considers privacy preservation. Many of these advanced classification methods are based on supervised machine learning approaches, and this will make it difficult to integrate them into a privacy-preserving framework, because the party undertaking the matching in such a case usually does not have access to the original, un-encoded data values that are used for the matching. Any classifier that requires (manually prepared) training data thus becomes cumbersome, if not impossible. Unsupervised classification techniques [15, 46] are thus required that do not rely upon the original, un-encoded data values.

Related to the need for automated classification techniques for record pairs is the challenge of how to assess the quality of the resulting matched records within a privacy-preserving framework. If the two data holders do not reveal any of their original, un-encoded record values to each other, how can one be assured that no truly matched record pairs were missed (i.e., how can recall be measured)? And of the matched records, given the party undertaking the matching does not see the original, un-encoded record values, how can precision be measured (i.e., how many of the classified matches correspond to true matches)? How to assess the quality of the matched data [2] within a privacy-preserving framework needs to be investigated when automated record pair classification techniques are being developed.

- **Scalability to large databases and real-time matching**

Besides improved secure matching and automated record pair classification techniques, scalability to very large databases is one of the major challenges of current approaches to privacy-preserving data matching. The computational and communication overheads of most approaches developed so far prohibit the matching of large databases that contain many millions of

records. What is required are novel techniques that scale linearly with the number of records to be matched, and also techniques that take advantage of the parallel multi-core computing capabilities that increasingly become available on many modern computing platforms [47].

A related challenge is the capability to match very large databases with a stream of incoming query records in (near) real-time [48], because real-time matching in a privacy-preserving framework is becoming increasingly important. Examples include identity verification for credit card applications and matching of crime and terrorism databases for national security. Real-time geocoding is vital when data from a possible bio-terrorism attack needs to be geocoded to obtain the locations where victims are living and working, as in such situations it is crucial to know who has been in contact with a victim.

– Preventing re-identification

While privacy-preserving data and geocode matching assures that no data is being revealed to any of the parties involved in a matching project, or to any external attacker, the question still remains how the information about the matched records is being further handled. For example, even if only zipcode, gender and age values are being released to the researcher who is conducting a study using the matched data, then it will likely be possible for this researcher to use these three attributes and match them with other, external data that is publicly available, and then re-identify the people in the matched data set [26]. This obviously can lead to a loss of privacy and confidentiality for the individuals whose records have been re-identified.

Traditionally, work on methods that prevent re-identification has been done by statisticians [49], while recently computer scientists have also started to investigate this challenge [26, 50]. Privacy-preserving matching will only become practically relevant if it is combined with anonymisation techniques that can guarantee that no re-identification is possible in any case, even when the matched records would be further matched with additional data.

5 Conclusions and Research Directions

In this paper, an overview of data and geocode matching has been presented, the principal steps involved in data matching have been discussed, and the specific techniques used in geocode matching and reverse geocoding have been highlighted. Using several scenarios, the privacy and confidentiality issues that arise when data from different organisations is being matched or geocoded have been illustrated. An overview of the recently developed privacy-preserving matching and geocoding approaches has been provided, and the major challenges in privacy-preserving data and geocode matching have been discussed.

Future research direction should address the challenges described in Sect. 4.2, with the objective to make privacy-preserving data and geocode matching more practical. While partial solutions exist to all of the described challenges, to the best of the author’s knowledge no currently available privacy-preserving data matching approach is tackling all of them.

Specifically, the computational and communication overheads of current approaches should be reduced to allow matching of very large databases, and the currently available privacy-preserving matching approaches should be integrated with advanced machine learning based classification methods to enable automated and accurate matching. Finally, techniques that are specific to geocode matching should be developed, to allow privacy-preserving geocoding in situations where a data holder cannot geocode its data otherwise.

References

1. US Federal Geographic Data Committee. Homeland Security and Geographic Information Systems: How GIS and mapping technology can save lives and protect property in post-September 11th America. *Public Health GIS News and Information* (52) (May 2003) 21–23
2. Christen, P., Goiser, K.: Quality and complexity measures for data linkage and deduplication. In Guillet, F., Hamilton, H.J., eds.: *Quality Measures in Data Mining*. Volume 43 of *Studies in Computational Intelligence*., Springer (2007) 127–151
3. Winkler, W.E.: Overview of record linkage and current research directions. Technical Report RRS2006/02, US Bureau of the Census (2006)
4. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering* **19**(1) (2007) 1–16
5. Kelman, C.W., Bass, J.A., Holman, D.: Research use of linked health data – A best practice protocol. *ANZ Journal of Public Health* **26**(3) (2002) 251–255
6. Jonas, J., Harper, J.: Effective counterterrorism and the limited role of predictive data mining. *Policy Analysis* (584) (2006)
7. Wang, G., Chen, H., Xu, J.J., Atabakhsh, H.: Automatically detecting criminal identity deception: An adaptive detection algorithm. *IEEE Transactions on Systems, Man and Cybernetics (Part A)* **36**(5) (2006) 988–999
8. Bhattacharya, I., Getoor, L.: Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **1**(1) (2007)
9. Hernandez, M.A., Stolfo, S.J.: Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery* **2**(1) (1998) 9–37
10. Churches, T., Christen, P., Lim, K., Zhu, J.: Preparation of name and address data for record linkage using hidden Markov models. *BioMed Central Medical Informatics and Decision Making* **2**(9) (2002)
11. Baxter, R., Christen, P., Churches, T.: A comparison of fast blocking methods for record linkage. In: *ACM KDD Workshop on Data Cleaning, Record Linkage and Object Consolidation*, Washington DC (2003)
12. Christen, P.: Febrl – An open source data cleaning, deduplication and record linkage system with a graphical user interface. In: *ACM International Conference on Knowledge Discovery and Data Mining*, Las Vegas (2008) 1065–1068
13. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: *IJCAI Workshop on Information Integration on the Web*, Acapulco (2003) 73–78
14. Christen, P.: A comparison of personal name matching: Techniques and practical issues. In: *IEEE ICDM Workshop on Mining Complex Data*, Hong Kong (2006) 290–294

15. Christen, P.: Automatic record linkage using seeded nearest neighbour and support vector machine classification. In: ACM International Conference on Knowledge Discovery and Data Mining, Las Vegas (2008) 151–159
16. Clarke, D.: Practical introduction to record linkage for injury research. *Injury Prevention* **10** (2004) 186–191
17. Christen, P., Willmore, A., Churches, T.: A probabilistic geocoding system utilising a parcel based address file. In: AusDM'04, Cairns, Australia, Springer LNAI 3755 (2006) 130–145
18. Paull, D.: A geocoded national address file for Australia: The G-NAF what, why, who and when? PSMA Australia Limited, Griffith, ACT, Australia (2003) Available at: <http://www.g-naf.com.au/>.
19. Cayo, M.R., Talbot, T.O.: Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics* **2**(10) (2003)
20. Brownstein, J.S., Cassa, C., Kohane, I.S., Mandl, K.D.: Reverse geocoding: Concerns about patient confidentiality in the display of geospatial health data. *AMIA Annual Symposium Proceedings* **2005** (2005) 905
21. Brownstein, J.S., Cassa, C., Mandl, K.D.: No place to hide—reverse identification of patients from published maps. *New England Journal of Medicine* **355**(16) (2006) 1741–1742
22. Curtis, A.J., Mills, J.W., Leitner, M.: Spatial confidentiality and GIS: Re-engineering mortality locations from published maps about Hurricane Katrina. *International Journal of Health Geographics* **5**(1) (2006) 44–56
23. Australian Attorney-General's Department, Standing Committee of Attorney's-General: Model criminal law officers' committee: Final report on identity crime. Canberra (March 2008)
24. Chaytor, R., Brown, E., Wareham, T.: Privacy advisors for personal information management. In: SIGIR Workshop on Personal Information Management, Seattle, Washington (2006) 28–31
25. Fienberg, S.E.: Privacy and confidentiality in an e-Commerce world: Data mining, data warehousing, matching and disclosure limitation. *Statistical Science* **21**(2) (2006) 143–154
26. Sweeney, L.: K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10**(5) (2002) 557–570
27. Christen, P.: Privacy-preserving data linkage and geocoding: Current approaches and research directions. In: IEEE ICDM Workshop on Privacy Aspects of Data Mining, Hong Kong (2006) 497–501
28. Sweeney, L.: Privacy-enhanced linking. *ACM SIGKDD Explorations* **7**(2) (2005) 72–75
29. Christen, P., Churches, T.: Secure health data linkage and geocoding: Current approaches and research directions. In: National e-Health Privacy and Security Symposium, Brisbane, Australia (2006)
30. Wartell, J., McEwen, T.: Privacy in the information age: A guide for sharing crime maps and spatial data. Institute for Law and Justice, NCJ 188739 (July 2001)
31. Rushton, G., Armstrong, M.P., Gittler, J., Greene, B.R., Pavlik, C.E., West, M.M., Zimmerman, D.L.: Geocoding in cancer research – A review. *American Journal of Preventive Medicine* **30**(2S) (2006) 16–24
32. Rivest, R.L.: Chaffing and winnowing: Confidentiality without encryption. MIT Lab for Computer Science (1998) Available at: <http://theory.lcs.mit.edu/~rivest/chaffing.txt>.

33. Churches, T.: 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)
34. Bouzelat, H., Quantin, C., Dusserre, L.: Extraction and anonymity protocol of medical file. In: *AMIA Fall Symposium*. (1996) 323–327
35. Dusserre, L., Quantin, C., Bouzelat, H.: A one way public key cryptosystem for the linkage of nominal files in epidemiological studies. *Medinfo* **8**(644–7) (1995)
36. Quantin, C., Bouzelat, H., Allaert, F.A., Benhamiche, A.M., Faivre, J., Dusserre, L.: Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. *Methods of Information in Medicine* **37**(3) (1998) 271–277
37. Quantin, C., Bouzelat, H., Allaert, F.A., Benhamiche, A.M., Faivre, J., Dusserre, L.: 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) (1998) 117–122
38. Quantin, C., Bouzelat, H., Dusserre, L.: Irreversible encryption method by generation of polynomials. *Medical Informatics and the Internet in Medicine* **21**(2) (1996) 113–121
39. Schneier, B.: *Applied cryptography: Protocols, algorithms, and source code in C*, 2nd Edition. John Wiley & Sons, Inc., New York (1995)
40. Ravikumar, P., Cohen, W.W., Fienberg, S.E.: A secure protocol for computing string distance metrics. In: *IEEE ICDM Workshop on Privacy and Security Aspects of Data Mining*, Brighton, UK (2004)
41. Atallah, M.J., Kerschbaum, F., Du, W.: Secure and private sequence comparisons. In: *ACM Workshop on Privacy in the Electronic Society*, Washington DC (2003) 39–44
42. O’Keefe, C.M., Yung, M., Gu, L., Baxter, R.: Privacy-preserving data linkage protocols. In: *ACM Workshop on Privacy in the Electronic Society*, Washington DC (2004) 94–102
43. Churches, T., Christen, P.: Some methods for blindfolded record linkage. *BioMed Central Medical Informatics and Decision Making* **4**(9) (2004)
44. Al-Lawati, A., Lee, D., McDaniel, P.: Blocking-aware private record linkage. In: *International Workshop on Information Quality in Information Systems*, Baltimore (2005) 59–68
45. Inan, A., Kantarcioglu, M., Bertino, E., Scannapieco, M.: A hybrid approach to private record linkage. In: *IEEE International Conference Data Engineering*. (2008) 496–505
46. Christen, P.: Automatic training example selection for scalable unsupervised record linkage. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer LNAI 5012, Osaka, Japan (2008) 511–518
47. Guisado-Gamez, J., Prat-Perez, A., Nin, J., Munes-Mulero, V., Larriba-Pey, J.L.: Parallelizing record linkage for disclosure risk assessment. In: *Privacy in Statistical Databases*, Istanbul, Turkey, Springer LNCS 5262 (September 2008) 190–202
48. Christen, P., Gayler, R.: Towards scalable real-time entity resolution using a similarity-aware inverted index approach. In: *AusDM’08, CRPIT vol. 87*, Glenelg, Australia (2008) 51–60
49. Winkler, W.E.: Masking and re-identification methods for public-use microdata: Overview and research problems. In: *Privacy in Statistical Databases*, Barcelona, Spain, Springer LNCS 3050 (2004) 216–230
50. Malin, B., Sweeney, L.: A secure protocol to distribute unlinkable health data. In: *American Medical Informatics Association 2005 Annual Symposium*, Washington DC (2005) 485–489