Chapter 5
Advanced Record Linkage Methods and Privacy Aspects for Population Reconstruction—A Survey and Case Studies

Peter Christen, Dinusha Vatsalan and Zhichun Fu

Abstract Recent times have seen an increased interest into techniques that allow the linking of records across databases. The main challenges of record linkage are (1) scalability to the increasingly large databases common today; (2) accurate and efficient classification of compared records into matches and non-matches in the presence of variations and errors in the data; and (3) privacy issues that occur when the linking of records is based on sensitive personal information about individuals. The first challenge has been addressed by the development of scalable indexing techniques, the second through advanced classification techniques that either employ machine learning- or graph-based methods, and the third challenge is investigated by research into privacy-preserving record linkage (PPRL). In this chapter, we describe these major challenges of record linkage in the context of population reconstruction. We survey recent developments of advanced record linkage methods, discuss two real-world case studies, and provide directions for future research.

5.1 Introduction

In the past decade, record linkage has attracted much interest by researchers and practitioners from various domains, including national census, health and social science research, businesses, and crime and fraud detection (Christen 2012a; Herzog and Scheuren 2007; Naumann 2010; Talburt et al. 2011). Also known as

P. Christen (✉) · D. Vatsalan · Z. Fu
Research School of Computer Science, The Australian National University, Canberra ACT 0200, Australia
e-mail: peter.christen@anu.edu.au

D. Vatsalan
e-mail: dinusha.vatsalan@anu.edu.au

Z. Fu
e-mail: sally.fu@anu.edu.au

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_5

data linkage, entity resolution, data matching, or duplicate detection, these techniques aim to identify and link all records that refer to the same real-world entities within a single or across several databases. In most applications, the entities under consideration are people, such as customers or patients.

The two areas where record linkage has traditionally been employed are national censuses (Winkler 2006) and the health domain (Kelman et al. 2002; Newcombe 1988). Most record linkage systems in these areas are based on the probabilistic record linkage approach developed by Newcombe and Kennedy (1962) and formalised by Fellegi and Sunter in 1969.

More recently, computer scientists have developed various techniques that allow the linking or deduplication of large databases with the aim to, for example clean customer records (Hernandez and Stolfo 1995) or identify fraudsters and criminals in financial and national security databases (Jonas and Harper 2006). Record linkage and deduplication techniques are also being employed to remove duplicate entries returned by search engines (Su et al. 2009), or to identify all bibliographic records of by the same author in publication databases (Lee et al. 2007).

Social scientists working in the area of demographics and genealogy have also employed record linkage techniques, commonly using historical census, or birth, death, and marriage (BDM) data (Fure 2000; Newton 2013; Quass and Starkey 2003; Reid et al. 2002; Ruggles 2002). The aim of such linkages is to identify and link not just individuals across two or more databases, but rather to create complete family trees over significant periods of time (Antonie et al. 2014a; Bloothooft 1995; Fu et al. 2014a). Such reconstructed (or reconstituted) family trees allow social scientists to investigate many aspects of past societies, such as changes in employment, mobility, fertility and morbidity, and even the genetic factors of certain diseases (Glasson et al. 2008).

Compared to contemporary data, the major challenges specific to the linking of historical data, which are based on census returns or BDM registers, are:

- The generally low levels of literacy of both census collectors and householders meant census items were often not recorded correctly. Dates of birth, and even ages, were commonly not known, and addresses were not clearly defined. There were no standard classifications of employment categories.
- Over time people moved, died, and were born, and so the structure of households and families changed significantly. Even if census returns are available for a full country, immigration and emigration mean a significant number of individuals simply ‘appear’ or ‘disappear’ without birth or death records. The influence of people’s movements is significantly worsened if only a small subset of census returns, like from a certain district or area, is available for research.
- Both given- and surnames often had strong local distributions. It was not uncommon for a large portion of a population to have one of a few common names.
- Only a small number of attributes were collected in many national censuses in the nineteenth century. For each individual they usually included the name, age, gender, relationship to the head of household, and occupation. Other data

sources, such as vital and parish registers (containing birth, baptism, death, and marriage records), can also provide rich sources of detailed information about families and their structures (Newton 2013; Reid et al. 2002).
- Historical documents are commonly hand-written and therefore have to be scanned and transcribed, either manually or automatically using optical character recognition techniques. These processes are likely to introduce further errors and variations into the data (Block and Star 1995).

Contemporary administrative and census databases are increasingly used for social science research. While present-day data are generally of higher quality and contain more detailed information, they pose their own set of challenges:

- As more information is being collected, today's databases not only become larger but they also contain more details about individuals, and they might also contain more complex types of data (such as text or multimedia documents). Linking very large databases poses significant computational challenges, as will be discussed in Sect. 5.2.
- The data collected are about people who are still alive, and therefore can contain sensitive information, for example about a person's health or their financial details. In today's 'Big Data' society, such information is highly valuable for organisations such as advertisers, insurers, financial institutions, and even governments, because it can facilitate for example specific individual targeting of advertisements, or the calculation of highly predictive credit risk scores (Siegel 2013). Privacy and confidentiality are especially of concern when records are linked across databases held by different organisations, as we will discuss in Sect. 5.3.

This chapter extends an earlier shorter workshop paper on the same topic (Christen 2014). In the following section, we provide a brief overview of advanced methods and techniques that have been developed in recent years. In Sect. 5.3 we discuss privacy issues relevant to record linkage and we summarise techniques that have been developed to facilitate linking databases across organisations without the need to reveal private or confidential information. In Sect. 5.4 we then illustrate, using two case studies, the issues discussed in Sects. 5.2 and 5.3. In Sect. 5.5 we present our view of important research directions for record linkage in the context of population reconstruction. We conclude this chapter in Sect. 5.6 with a summary of our findings. We also provide an extensive bibliography to relevant work.

5.2 Advanced Record Linkage Methods

A variety of techniques have been developed that allow the linking of large databases. The main areas of research have been to improve scalability to linking large databases, and to improve linkage quality using advanced classification techniques.

5.2.1 Scalable Indexing Techniques

When two databases are linked, each record from one database potentially has to be compared with all records from the other database. The vast majority of these comparisons will be between records that are not matches (i.e. refer to different entities). Indexing is the process of reducing this possibly very large number of record pairs that need to be compared in detail between databases by splitting each database into smaller sets of blocks or clusters, or by sorting the databases. The aim is to identify *candidate record pairs* from records in the same blocks or clusters that likely correspond to true matches, and that need to be compared in detail, generally using approximate string comparison functions (Christen 2012a).

The traditional blocking approach employs a *blocking criteria* (a single or set of attributes) to insert each record into one block (Fellegi and Sunter 1969). For example, if a 'postcode' attribute is used as blocking criteria then all records with postcode '2000' are inserted into the same block. Only records within the same block are then compared with each other. The sorted neighbourhood approach (Hernandez and Stolfo 1995) sorts a database according to *sorting criteria* (usually a set of concatenated attributes) and then moves a sliding window over the sorted database. Only records that are within a certain window are compared with each other.

Many of the recently developed indexing techniques insert each record into more than one block, thereby aiming to overcome errors in attribute values (Christen 2012b). Overlapping clusters (called canopies), sorted suffix arrays, and q-gram-based indexing, are examples of such techniques. A different approach is to map records into a multi-dimensional space such that the distances between records are preserved (Jin et al. 2003). A multi-dimensional index data structure together with nearest-neighbour queries are then used to extract blocks of candidate records.

Adaptive techniques that, based on the characteristics of the data, dynamically modify the size of the window in the sorted neighbourhood method (Draisbach et al. 2012; Yan et al. 2007) or in suffix array-based indexing (de Vries et al. 2011) have recently shown to obtain blocks of higher quality. Other recent work has investigated indexing techniques for real-time record linkage, where a stream of query records is to be linked in sub-second time to a database of entity records (Christen et al. 2009; Ioannou et al. 2010; Ramadan et al. 2014). Related to real-time record linkage are approaches that allow for dynamic databases, where records are added, modified, or removed, on an ongoing basis (Dey et al. 2010; Ioannou et al. 2010).

While traditional indexing approaches require manual decisions about the choice of blocking criteria, several approaches have been proposed to learn optimal blocking criteria either using training data (pairs or groups of records known to be true matches or non-matches) (Bilenko et al. 2006; Michelson and Knoblock 2006), or more recently by exploring the distribution of attribute values in records and the similarities between them (Kejriwal and Miranker 2013). The aim of such learning techniques is to find blocking criteria that lead to small blocks which contain mostly

matches only and overall have a high coverage of all matches (if they are known from training data). Generally, as will be discussed next, techniques that make use of true matches and non-matches obtain blocking results of higher quality compared to techniques that are only based on data distributions.

Only limited experimental evaluations have been conducted to compare the performance of indexing techniques. Christen (2012b) identified that none of 12 variations of six techniques outperformed all others when employed on several data sets, and that one of the most important factors for efficient and accurate indexing is the definition of an appropriate blocking criteria.

None of the indexing techniques discussed here is specific to a certain type of data, and therefore any can be used in the context of linking data for population reconstruction. However, given the often low quality especially of historical data, techniques should be applied that are able to cope with 'dirty' data and bring matching records together that likely contain errors and variations. To this end, techniques that insert each record into several blocks can be of advantage (at the cost of having to compare a larger number of candidate pairs), as can be techniques that incorporate domain expertise to guide the indexing process [for example by learning good blocking criteria (Bilenko et al. 2006; Kejriwal and Miranker 2013; Michelson and Knoblock 2006)].

5.2.2 Accurate Classification Techniques

The objective of record linkage classification is to decide if a pair or group of records is a *match* (assumed to refer to the same real-world entity) or a *non-match* (refer to different entities). In the traditional probabilistic record linkage approach (Fellegi and Sunter 1969), each compared record pair is classified independently into one of three classes (*matches*, *non-matches* and *potential matches*). The third class is those pairs or groups of records that require manual classification through a clerical review process (Christen 2012a).

Besides requiring an often time consuming manual clerical review step, this traditional approach has several other drawbacks. First, it assumes independence between attributes. Statisticians have investigated approaches that allow dependencies between some attributes to be modelled (Winkler 2006), and have achieved improved classification outcomes in some situations. Second, the estimation of the parameters needed for the probabilistic record linkage approach is a non-trivial undertaking and requires knowledge about the error rates in the databases to be linked (which is often difficult to obtain) (Herzog et al. 2007). Third, individual pair-wise classification can lead to a violation of the transitive closure property (if record pairs (a, b) and (a, c) are classified as matches, then pair (b, c) must also be a match).

Machine learning based approaches aim to overcome these deficiencies. They are either following a supervised learning approach, where training data in the form of known matching and non-matching record pairs are required (Elmagarmid et al.

2007), or they are based on unsupervised clustering techniques which group records according to their similarities (Naumann and Herschel 2010). While supervised approaches generally achieve higher linkage quality, their main drawback is the challenge of obtaining a large number of suitable training examples. Active learning techniques aim to overcome this drawback (Arasu et al. 2010; Bellare et al. 2012). They select a small number of difficult to classify record pairs and present these to a domain expert for manual classification, followed by a re-training of the classification model. This process is repeated until high enough linkage quality is obtained.

Several collective classification techniques for record linkage have recently been developed. Compared to the traditional classification of individual record pairs, based on a graph representation of the databases to be linked these techniques aim to find an overall optimal solution when assigning records to entities. Both Bhattacharya and Getoor (2007) and Kalashnikov and Mehrotra (2006) build a graph with records as nodes and relational and attribute similarities between them as edges. On the other hand, Dong et al. (2005) build a dependency graph where each attribute value pair is represented as a node that contains the similarity between the two values. An overall optimal classification is calculated in an unsupervised way by iteratively merging or splitting parts in such a graph into smaller sub-graphs, such that at the end of the process each sub-graph corresponds to an entity. A related technique is group linkage (On et al. 2007), where groups rather than individual records are considered and linked based on some form of group similarity.

Most experimental evaluations of these collective and group linkage techniques have been conducted using bibliographic databases, where different types of entities (authors, papers, venues, and affiliations) provide a rich and well-defined setting of relational information between entities. Compared to historical data, the quality of bibliographic data is generally high, but ambiguities occur, for example when non-standardised abbreviations of conferences or journals are recorded, only the initials of authors are given, or several authors have the same name and even work in the same research area. For two ambiguous author records, co-author similarities or having published in similar journals or conferences can provide the evidence needed to decide if the two records refer to the same author or not. The databases used to evaluate collective classification techniques generally contained less than one million records, and scalability of these techniques to very large databases has only been investigated recently (Rastogi et al. 2011).

Only limited work has been conducted in machine learning-based record linkage for population reconstruction. Antonie et al. (2014a, b) use a support vector machine classifier to link historical Canadian census data, while Efremova et al. (2015) use a linear scoring model to weight different similarity measures in the context of matching historical Dutch BDM records. These works highlight the successful application of supervised classification techniques for population reconstruction, but they also discuss the challenges in acquiring the required training data.

Fu et al. (2014a, 2011b, 2012) have recently investigated group linkage methods on historical census data by treating households as groups and combining pair-wise

record linkage with household linkage. Their evaluation on UK census data showed a significant reduction in the number of multiple links (i.e. where a single record from one database is linked to several records in another database).

The unique structure between records within a family or household has only recently been explored for record linkage. While most personal details of people change over time, some aspects of the relationships between the members of a family or household keep constant even over long periods of time. For example, the age differences between two parents, and between parents and their children, do not change (assuming they are recorded accurately). As we will illustrate in Sect. 5.4.1, Fu et al. (2014b) recently proposed to build one graph per household using such time-invariant information as edge attributes, and they showed that such an approach can help to improve household matching in historical census data. Graph-based approaches can exploit such rich sources of structural information and allow the development of improved record linkage techniques in the context of population reconstruction.

5.3 Privacy Aspects in Record Linkage

Due to the lack of unique entity identifiers, record linkage is generally based on comparing partially identifying personal details of individuals, such as their names, addresses, dates of birth, and so on. When historical data are being linked then usually no privacy concerns are being raised, because these data do not contain any information about living individuals. However, as social science research increasingly requires the linking of contemporary databases obtained from diverse sources, privacy and confidentiality issues become crucially important. National census agencies are currently considering the use of anonymisation techniques to facilitate matching their databases with records sourced from public as well as private administrative data (Office for National Statistics 2013).

While a single database that contains the personal details of individuals can already contain sensitive information, linking records sourced for instance from government agencies with records from commercial databases can reveal information that is highly sensitive. For example, an individual's social security (unemployment) record linked with their financial details obtained from a bank database would be of high value for a credit rating agency. As recent events in the context of national security data leakages have shown (Edward Snowden's copying and releasing of thousands of top secret US government documents) (Toxen 2014), people are wary that their information is being collected by and shared across different organisations, especially if this is done by governments.

The linking of contemporary databases from diverse sources can allow studies at levels of detail and at scales otherwise not possible, and therefore safeguards must be in place to make sure no private or confidential information can be revealed. In the health domain, specific protocols (Churches 2003; Kelman et al. 2002) have been developed and are in use that split sensitive health data from the attributes

used for the actual linkage. These protocols, however, still require a trusted third party to conduct the linkage using the actual personal details of individuals. Ideally, it should be possible to conduct record linkage without the need of any sensitive information to be exchanged between the parties that are involved in a record linkage project.

Researchers working in the area of 'privacy-preserving record linkage' (PPRL) are aiming to achieve this goal (Verykios and Christen 2013). Vatsalan et al. (2013) provide an extensive review and propose a taxonomy of current PPRL techniques, and they discuss research challenges and directions. The basic ideas of PPRL techniques are to (somehow) encode (or mask) the databases at their sources and to conduct the linkage using only these encoded data (i.e. no sensitive data are ever exchanged between parties). At the end of such a PPRL process, the database owners only learn which of their own records have a high similarity with certain records from the other database(s). The database owners can then negotiate the next steps, such as exchanging the values in certain attributes of the linked records, or sending selected attribute values to a third party (for example a researcher, as discussed in Sect. 5.4.2).

The two basic scenarios in PPRL are two- and three-party protocols. In the latter type, a linkage unit is conducting the actual linkage based on encoded data received from the two database owners. On the other hand, in two-party protocols the two database owners directly exchange encoded data between them. The advantage of two-party over three-party protocols is that they are more secure, as there is no possibility of collusion between one of the database owners and the linkage unit. However, two-party protocols are generally more complex in order to make sure that the two database owners cannot infer any sensitive information from each other during the PPRL process.

Research into PPRL started in the mid 1990s, and the developed techniques can be categorised into three generations (Vatsalan et al. 2013). The first only considered the exact matching of attribute values without revealing these values. These techniques basically convert attribute values into hash codes (bit-patterns of a certain length) using one-way hash algorithms such as SHA or MD5 (Schneier 1996), and then compare the generated hash codes in an exact fashion. These hash codes are secure in that having only access to a hash code makes it nearly impossible (with current computing techniques) to find the corresponding plain-text string in a reasonable amount of time. The major drawback of the first generation of PPRL techniques is that even a single character difference between attribute values results in completely different hash codes, and so only exact matching of values is possible. As data, especially personal details such as names and addresses, often contain variations and errors, exact matching does not work well in most practical linkage situations.

The second generation of PPRL techniques aimed to overcome this drawback by allowing for approximate matching. Approaches for secure edit-distance, Jaccard and overlap similarity, and Cosine distance have been developed, with several recent surveys providing comparative evaluations of such techniques (Durham et al. 2012; Karakasidis and Verykios 2010; Trepetin 2008; Vatsalan et al. 2013;

Verykios et al. 2009). A variety of techniques have been investigated, including Bloom filters (bit-arrays) (Schnell et al. 2009; Vatsalan and Christen 2012), phonetic encoding (such as Soundex or NYSIIS) (Karakasidis and Verykios 2009), random and public reference values (Karakasidis et al. 2011; Pang et al. 2009; Vatsalan et al. 2011), embedding spaces (into multi-dimensional spaces) (Scannapieco et al. 2007; Yakout and Atallah 2009), and secure multi-party computation (Atallah et al. 2003; Inan et al. 2008; Li et al. 2011; Ravikumar et al. 2004). The interested reader is referred to the above cited survey articles for details.

While allowing for approximate matching was a significant improvement for PPRL, the problem of scalability to linking large databases has only recently been considered in the third generation of PPRL techniques (Al-Lawati et al. 2005; Bonomi et al. 2012; Durham 2012; Inan et al. 2010; Karakasidis 2012; Karapiperis and Verykios 2014; Kuzu et al. 2013; Sehili et al. 2015; Vatsalan et al. 2013a). Different techniques have again been developed which combine traditional indexing techniques (Christen 2012b) with encoding, perturbation, or cryptographic approaches (Vatsalan et al. 2013). Thus far, only a few small comparative studies of such techniques have been published (Durham 2012; Vatsalan et al. 2013a, 2014). The issues involved in evaluating PPRL techniques have also received increased attention in recent times (Vatsalan et al. 2014).

5.4 Case Studies

In this section we present two case studies with a focus on advanced record linkage techniques being employed to population reconstruction. The first study discusses the use of group and graph linking in the context of linking historical census data, while the second discusses approaches to preserving privacy when linking contemporary data from several sources from both private and public organisations.

5.4.1 Advanced Linking of Historical UK Census Data

Our case study uses historical census returns collected from the district of Rawtenstall, which in the nineteenth century was a small cotton textile manufacturing town in North-East Lancashire in the United Kingdom (UK). Currently released historical census data in this area were collected since 1851 in ten-year intervals. The original data were hand-filled census forms, which contain 12 attributes, that for each individual residing in a household include the address, full name, age, gender, occupation, place of birth, and their relationship to the head of the household.

These hand-filled census forms were transcribed manually onto enumerator's returns sheets, and these sheets were subsequently scanned into digital form. Since the late 1990s, various organisations began transcribing these data from images into

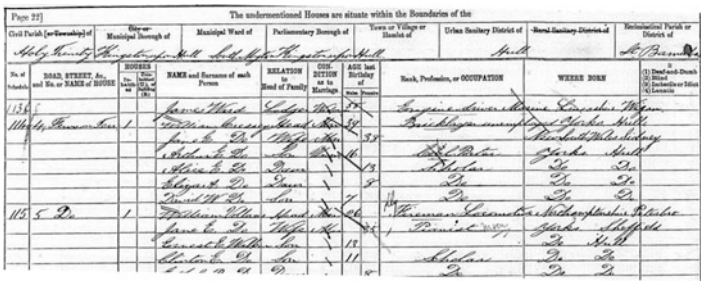


Fig. 5.1 Historical census sample

tabular form and stored them in spreadsheets where they could be examined by members of the public. A sample of a scanned image is shown in Fig. 5.1. Our collection consists of six data sets, with around 160,000 records in total, corresponding to the censuses from 1851 to 1901.

To link such historical census data, several key steps are necessary to calculate the similarities between records from the individual data sets (Christen 2012b). These steps include data cleaning and standardisation to improve data quality and make attribute values more consistent before comparison; blocking or indexing, as discussed in Sect. 5.2.1, to subdivide a data set into blocks so that records in a block are only being compared with other records in the same block in the comparison step; and finally the classification of the compared record pairs into matches and non-matches, as was discussed in Sect. 5.2.2.

The differences between traditional record linking methods and those based on group or graph methods are in the final classification step. Traditional approaches only perform linkage at the record-pair level, relying only on the output of record attribute similarities to classify record pairs. In practice, this strategy often faces difficulties because most historical data have significant data quality problems, and only limited details about people are available in historical (census) data that can be compared attribute-wise between records. Group- or graph-based methods, on the contrary, consider households (or families) as integral entities, and use the whole of household information to improve the effectiveness and accuracy of record linkage.

To illustrate how household information can help group- and graph-based record linkage, let us consider the following example. Table 5.1 shows a household with four people, consisting of the parents and two children, extracted from the 1871

Table 5.1 1871 household sample. The example record for *Sarah Ashworth* is highlighted in italics

ID	Address	Surname	Firstname	Relation_to_head	Sex	Age
25531	union street	ashworth	john	head	m	30
25532	union street	ashworth	alice	wife	f	28
25533	union street	ashworth	richard	son	m	4
25534	*union street*	*ashworth*	*sarah*	*daughter*	*f*	*2*

census return. The key attributes and their values for each member are displayed, with 'ID' being a unique record identifier.

When applying traditional pair-wise record linkage between 1871 and 1881 census returns, we can see that Sarah Ashworth (ID 25534) has two matched records in 1881. One (ID 12534 in Table 5.2) lived in the same address but with a wrong age, and the other (ID 20858 in Table 5.3) lived at a different address with the correct age. Based only on the attributes in these records, it is difficult to determine which is the correct Sarah. Most pair-wise record linkage approaches will take the match with ID 12534 to be the correct one because it has a higher similarity to the 1871 Sarah than the second option, because street addresses normally contain more distinguishing information than age. As example, the pair-wise linking method by Fu et al. (2014a) uses approximate string comparison functions (Christen 2012a) on the address and name attributes and absolute differences on the age attribute, and gives a total similarity score of 0.9 for the record pair with ID 25534 and ID 12534 of Sarah Ashworth, higher than the total similarity score between ID 25534 and ID 20858, which is 0.84. This shows that pair-wise record linkage is not always reliable.

If we take other household members into consideration, it is obvious that record with ID 20858 in Table 5.3 should be the true match for record with ID 25534 in Table 5.1. The reason is clear: the names and ages of Sarah's parents and her brother Richard in this 1881 household (with Richard abbreviated as 'rd' in 1881) also match the corresponding members in the 1871 household, while the other members of the household in Table 5.2 do not. Therefore, household information can greatly help the decision making so as to reduce the ambiguity that arises from the pair-wise linkage results.

Table 5.2 Wrongly matched 1881 household

ID	Address	Surname	Firstname	Relation_to_head	Sex	Age
12532	union street	ashworth	henry	head	m	48
12533	union street	ashworth	eruble	wife	f	47
12534	*union street*	*ashworth*	*sarah*	*daughter*	*f*	*18*
12535	union street	ashworth	john	son	m	12

Table 5.3 Correctly matched 1881 household

ID	Address	Surname	Firstname	Relation_to_head	Sex	Age
20855	whittle st	ashworth	john	head	m	40
20856	whittle st	ashworth	alice	wife	f	36
20857	whittle st	ashworth	rd.	son	m	14
20858	*whittle st*	*ashworth*	*sarah*	*daughter*	*f*	*12*
20859	whittle st	ashworth	john	son	m	8
20860	whittle st	ashworth	harold	son	m	3

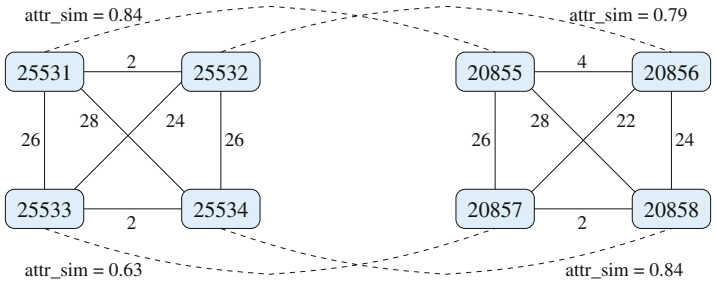
The key to utilising household information is how to model household members and their relationship. Group and graph linking are two methods aiming to solve this problem. Group linking (Fu et al. 2011b, 2012, 2014a; On et al. 2007) generates group similarity scores for each pair of households. Such household pair similarities are calculated in several steps. First, the number of household members in each household is counted. Then, the sum of the pair-wise record similarity scores between the household pairs is calculated. This sum is normalised by the number of distinct members in the two households being compared so as to generate the household similarity score. When two or more households are compared with a target household, the one with the highest household similarity is being matched.

For example, we know that the households in Tables 5.1 and 5.2 both have four members. Pair-wise linking results, again following the approach taken by Fu et al. (2014a), show that the similarity between records with ID 25531 and ID 12535 is 0.9, between ID 25532 and ID 12533 is 0.66, and between ID 25534 and 12534 is 0.9. Then, the group linking similarity score, using the bipartite similarity (Fu et al. 2011b), between this household pair is calculated as $(0.66 + 0.9 + 0.9)/(4 + 4 - 3) = 0.49$, i.e., the sum of record pair-wise similarities divided by the number of distinct members in these two households. The same calculation gives a group linking similarity score between the two households in Tables 5.1 and 5.3 of 0.78. Based on this it becomes clear that the households in Tables 5.1 and 5.3 are matched, and that Sarah Ashworth with ID 25534 and ID 20858 are matches.

Graph-based linking can be considered as an extension of the group linking step. Graph linking does not only consider the similarity of all record pairs in two households, it also takes structural information of households into consideration. While personal information, such as marital status, address and occupation, may change over time, surnames of women may change after marriage, and even ages may change due to different times of the year for census collection or input errors, some aspects of the relationships between household members generally remain unchanged. Such relationship aspects include, but are not limited to, age and generation difference, and role-pairs of two individuals in a household (Fu et al. 2014b). By incorporating such relationship aspects between household members into the linking model the linking accuracy can be improved.

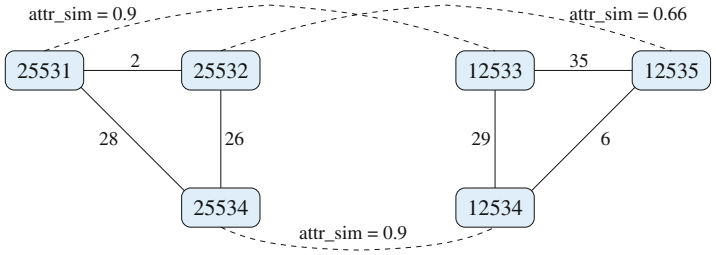
The graph method in (Fu et al. 2014b) treats members in a household as the vertices (nodes) of a graph, and uses edges to show the relational aspects between these vertices. The method first calculates record-pair similarities, which are used to find matched candidate record pairs with high similarities. These pairs are then used to connect the graphs of two candidate households. This transforms the household linking problem into a graph matching problem. The graph similarity score for each pair of households is calculated as the weighted sum of the vertex and edge similarities.

As an example of graph-based linkage using the three households from Tables 5.1, 5.2 and 5.3, Fig. 5.2 shows the graph generated between the 1871 and the correctly matched household, while Fig. 5.3 shows the graph generated between the 1871 and the wrongly matched household. Only the AGE attribute is used in



$$avr\_node\_sim = (0.84 + 0.79 + 0.63 + 0.84)/4 = 0.78$$
$$avr\_edge\_sim = (0.50 + 1.00 + 1.00 + 0.92 + 0.92 + 1.00)/6 = 0.89$$
$$household\_sim = \lambda \times avr\_node\_sim + (1 - \lambda) \times avr\_edge\_sim = \mathbf{0.84}$$

Fig. 5.2 Graph structure and similarity calculations of the two matched households from Tables 5.1 and 5.3. Edge values are absolute differences in AGE values, while the dotted lines show attribute similarities between records in the two households. We set the weighting parameter (Fu et al. 2014b) $\lambda = 0.5$



$$avr\_node\_sim = (0.90 + 0.66 + 0.90)/3 = 0.82$$
$$avr\_edge\_sim = (0.06 + 0.97 + 0.23)/3 = 0.42$$
$$household\_sim = \lambda \times avr\_node\_sim + (1 - \lambda) \times avr\_edge\_sim = \mathbf{0.62}$$

Fig. 5.3 Graph structure and similarity calculations of the two non-matched households from Tables 5.1 and 5.2. We again set the weighting parameter $\lambda = 0.5$

this example (in practice, the relationships between individuals would also be used). Only records that have high attribute similarities between the households are included in these graphs. The shown edge attribute values are age differences between records in the corresponding vertices, while the dotted lines between vertices in these household graphs correspond to the record attribute similarities calculated in the pair-wise linkage step. The edge (AGE) similarities are calculated as $age\_sim = 1.0 - abs(age\_diff)/max\_age$ (Christen 2012a).

When only the node similarities are considered, the similarity for the matched household pair is 0.78 for the households from Tables 5.1 and 5.3, which is lower than the similarity for the non-matched household pair (0.82), as shown in the

calculations in Figs. 5.2 and 5.3, respectively. On the other hand, when the age relationships between household members are considered, a higher overall similarity is calculated for the matched household pair (0.84) compared to the non-matched pair (0.62), resulting in correctly matched households.

The results on the six Rawtenstall data sets show that the proposed methods significantly reduce the number of multiple record and household matches, with a more than 85 % reduction using either group or graph linking approaches (Fu et al. 2011b, 2014b).

5.4.2 *Privacy-Preserving Record Linkage Across Several Organisations*

Linking records across several databases held by different organisations, using the common identifiers that contain personal information, often involves privacy and confidentiality concerns of the individuals represented by the records in these databases (Vatsalan et al. 2013b). Generally, organisations are not allowed or willing to exchange such personal and sensitive information due to privacy and confidentiality concerns as well as government or business regulations.

As an example scenario, assume a demographer who aims to investigate how mortgage stress (having to pay large sums of money on a regular basis to repay a house) is affecting people from different ethnic backgrounds, and with different education and employment levels, with regard to their mental and physical health. This research will require data from financial institutions, as well as different government agencies (social security, health, and eduction), and potentially other private sector providers (such as health insurers). Neither of these parties is likely willing or allowed by law to provide their databases to the researcher. The researcher only requires access to some attributes of the records that are linked across all these databases, but not the actual identities of the individuals that were linked. However, personal details are needed to conduct the actual linkage due to the absence of unique identifiers across all the databases. As was discussed in Sect. 5.3, PPRL aims to address this problem.

Assume three databases from three different organisations, as shown in Tables 5.4, 5.5 and 5.6, need to be linked in order to identify the matching entities across these databases. A set of common personal identifiers, which are first_name / given_name, last_name /surname, and postcode, are used as quasi-identifiers (QIDs) for conducting the linkage. Exchanging the actual values of these QIDs is not possible in this scenario as it would compromise the privacy and confidentiality of the individuals represented in these databases. Therefore, the linkage has to be conducted on masked (encoded) versions of the QID values which have a specific functional relationship with the actual QID values (Vatsalan et al. 2014).

There have been various masking functions proposed in the literature, as reviewed in Vatsalan et al. 2013b. Bloom filter encoding is one masking approach

Table 5.4 Example bank database

ID	First_name	Last_name	DOB	Gender	Postcode	Loan_type	Period	Amount	Paid
6723	peter	robert	20.06.72	M	2617	Mortgage	20	350,000	130,000
8345	miller	roberts	11.10.79	M	2602	Personal	5	10,000	1,900
9241	amelia	millar	06.01.74	F	2415	Mortgage	30	475,000	154,250

Table 5.5 Example social security database

SSN	Title	Last_name	First_name	Age	Postcode	Employment	Income	Benefits	Payment
490814	Mrs	amilia	smith	39	2642	teacher	60,000	child care	45,000
581233	Mr	peter	roberts	42	2617	engineer	110,000	family tax	50,000
932389	Mr	william	smith	69	3205	retired	–	pension	35,000

Table 5.6 Example health database

PID	Surname	Given_name	Age	Postcode	Sex	Pressure	Stress	Last_visited	Reason_of_visit
P1209	robertt	peter	41	2617	m	140/90	high	25 days ago	chest pain
P4204	miller	amelia	39	2415	f	120/80	high	61 days ago	headache
P4894	sieman	jeff	30	2602	m	110/80	normal	15 days ago	checkup

that has widely been used in PPRL (Durham 2012; Ranbaduge et al. 2014; Schnell et al. 2009; Sehili et al. 2015; Vatsalan and Christen 2012, 2014). An example of Bloom filter encoding is illustrated in Fig. 5.4. The Bloom filter encoded QID values can then be compared using a set-based similarity function such as the Dice-coefficient (Schnell et al. 2009). The Dice-coefficient of P Bloom filters $(b_1, \ldots, b_P)$ is calculated as:

$$dice\_sim(b_1, \ldots, b_P) = \frac{P \times c}{\sum_{i=1}^{P} x_i} \tag{5.1}$$

where c is the number of common bit positions that are set to 1 in all P Bloom filters (common 1-bits), and x_i is the number of bit positions set to 1 in b_i (1-bits), $1 \leq i \leq P$.

As discussed in Sect. 5.2.1, comparing all pairs or sets of records is not scalable due to the resulting quadratic or exponential complexities, respectively (Vatsalan et al. 2013b). Generally, private blocking or indexing techniques (Durham 2012; Karakasidis 2012; Kuzu et al. 2013; Ranbaduge et al. 2014; Vatsalan et al. 2013a)

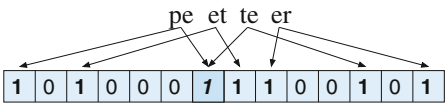


Fig. 5.4 Example Bloom filter encoding of value 'peter'. The q-grams ($q = 2$) of 'peter' are hash-mapped into a Bloom filter of $l = 14$ bits using $k = 2$ hash functions

are used to reduce the number of comparisons that are required in PPRL. For example, applying Soundex-based phonetic blocking (Christen 2006) on the three example databases using the surname/last_name attribute as the blocking criteria results in blocks as shown in Tables 5.7, 5.8 and 5.9 with their encoded Bloom filters (made-up) using first_name/given_name, last_name/surname, and postcode as the QIDs.

Records are then compared with only the records from other databases that are in the same block. In the running example, comparing records (Bloom filters) in the blocking key (Soundex code) 'r163' using the Dice-coefficient similarity function and classifying records as matches that have a *dice_sim* (Eq. 5.1) of at least a minimum threshold $s_t = 0.8$, are shown in Figure 5.5. The records with ID 6723 from Table 5.4, SSN 581233 from Table 5.5, and PID P1209 from Table 5.6 are classified as corresponding to the same person as the similarity of these (masked) records is $dice\_sim = 0.86$ (≥ 0.8). Identifying matching records from subsets of databases (e.g. ID 9241 from Table 5.4 and PID P4204 from Table 5.6) is also an important problem in PPRL which requires further research.

Schnell et al. (2009) and Durham (2012) proposed to use a third party (linkage unit) to compare and classify the Bloom filters from two database owners.

Table 5.7 Records in the example bank database (Table 5.4) with their blocks, QIDs, and Bloom filter encodings. The records in block 'r163' are highlighted in italics

Block	Rec_ID	QID	Bloom filter
r163	*6723*	*peter,robert,2617*	*1 1 1 1 0 1 0 1 0 0 1 1 0 1 1 0 1 0 1 1*
r163	*8345*	*miller,roberts,2602*	*1 1 1 0 0 0 1 0 1 0 1 0 1 1 1 0 1 1 0 1*
m460	9241	amelia,millar,2415	1 1 0 0 1 0 1 1 0 0 0 1 1 0 0 1 1 0 0 1

Table 5.8 Records in the example social security database (Table 5.5) with their blocks, QIDs, and Bloom filter encodings. The record in block 'r163' is highlighted in italics

Block	Rec_ID	QID	Bloom filter
s530	490814	amilia,smith,2642	1 1 0 1 0 1 1 0 0 1 1 0 1 0 0 1 1 0 1 0
r163	*581233*	*peter,roberts,2617*	*1 1 1 1 0 1 0 1 1 0 1 1 1 0 1 0 1 0 1 1*
s530	932389	william,smith,3205	1 0 0 1 0 1 1 0 0 1 1 1 0 0 1 1 0 1 0 0

Table 5.9 Records in the example health database (Table 5.6) with their blocks, QIDs, and Bloom filter encodings. The record in block 'r163' is highlighted in italics

Block	Rec_ID	QID	Bloom filter
r163	*P1209*	*peter,robertt,2617*	*1 1 1 1 0 1 0 1 0 1 1 1 0 1 1 1 1 0 1 1*
m460	P4204	amelia,miller,2415	1 1 0 1 1 0 1 1 0 0 0 1 1 0 0 1 1 1 0 1
s550	P4894	jeff,sieman,2602	0 1 0 1 1 0 0 0 1 0 1 1 0 0 1 0 0 1 0 1

	Candidate set	Bloom filters	x and c	Similarity	Class
1.	(6723, 581233, P1209)	1 1 1 1 0 1 0 1 0 0 1 1 0 1 1 0 1 0 1 1 1 1 1 1 0 1 0 1 1 0 1 1 1 0 1 0 1 0 1 1 1 1 1 1 0 1 0 1 0 1 1 1 0 1 1 1 1 0 1 1 & 1 1 1 1 0 1 0 1 0 0 1 1 0 0 1 0 1 0 1 1	$x_1 = 13$ $x_2 = 14$ $x_3 = 15$ $c = 12$	3×12 $(13 + 14 + 15)$ $= 0.86$	Match
2.	(8345, 581233, P1209)	1 1 1 0 0 0 1 0 1 0 1 0 1 1 1 0 1 1 0 1 1 1 1 1 0 1 0 1 1 0 1 1 1 0 1 0 1 0 1 1 1 1 1 1 0 1 0 1 0 1 1 1 0 1 1 1 1 0 1 1 & 1 1 1 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 1	$x_1 = 12$ $x_2 = 14$ $x_3 = 15$ $c = 7$	3×7 $(12 + 14 + 15)$ $= 0.51$	Non-match

Fig. 5.5 Comparison and classification of Bloom filters in block 'r163' from Tables 5.7, 5.8 and 5.9

A privacy risk with using a linkage unit is the possible collusion between a party and the linkage unit with the aim to learn about data from the other parties (Vatsalan et al. 2013b). A two-party protocol (Vatsalan and Christen 2012) was later proposed where the database owners iteratively exchange selected bits from their Bloom filters and classify the record pairs without requiring a third party. Most work in PPRL so far support the linkage of two sources only. However, two novel approaches for multi-party PPRL for more than two databases based on Bloom filter encodings were recently proposed (Ranbaduge et al. 2014; Vatsalan and Christen 2014). One of the main challenges with multiple parties is the exponential increase in the number of record sets that potentially have to be compared.

In addition to Bloom filter encoding, several other masking functions, ranging from computationally expensive cryptographic techniques (Lindell and Pinkas 2009) to differential privacy (Dwork 2006), *k*-anonymity (Sweeney 2002), reference values (Pang et al. 2009), and noise addition techniques (Karakasidis et al. 2011) have been used in the literature to preserve privacy while allowing the linkage. Other privacy components that need to be considered in a PPRL project are encrypted communication among the parties using public/private key pairs, secure generation and exchange of keys, employee confidentiality agreements to reduce internal threats, as well as secure connections and servers to reduce external threats.

5.5 Research Directions

Most advanced record linkage techniques have been developed by computer science researchers. The focus of these techniques was not only on data that contain personal information, as is generally required for population reconstruction, but often on bibliographic records, or consumer product or business data. Based on existing techniques and approaches, the following research directions can be identified:

- A main open challenge is how collective and graph-based classification techniques, that have shown to be highly accurate, can be used on personal data such as those available in (historical) census and BDM databases. Compared to the bibliographic databases on which such techniques so far have been evaluated, much less relational structure is available in personal data. Specifically, the number of different entity types, and their relationships, are more limited.
- Only limited work has been conducted on how to incorporate temporal information into the linkage process, such as personal details like name and address values that can change over time (Chiang et al. 2014; Christen and Gayler 2013; Li et al. 2011). However, such changes, especially in address attributes, occur regularly and at significant rates.
- As in many applications no or only a limited amount of training data in the form of true matches and non-matches are available, further investigating active learning techniques (Arasu et al. 2010; Bellare et al. 2012), specifically in the context of population reconstruction, could lead to significant reduction in the manual efforts currently required with traditional record linkage approaches. Furthermore, visualising, for example multiple households or families that were linked over time, and highlighting ambiguities and conflicts in the obtained linkages, could help to both better understand problems in linkage algorithms, and also improve the selection and preparation of manual training examples.
- Related to the previous point, given the generally low quality of historical data, developing (semi-) automatic data cleaning and standardisation techniques (Fu et al. 2011a), based on approaches that learn the characteristics of data errors and variations, will significantly reduce the time consuming and cumbersome process of manual data cleaning that is still commonly required today. The requirements of training data of such learning algorithms should be minimised, by for example employing active learning (Arasu et al. 2010; Bellare et al. 2012) or bootstrapping approaches where increasingly accurate models are trained in an iterative fashion (Churches et al. 2002). Additionally, such learning techniques should also be transferable from one domain to another, or allow re-training with little (manual) effort.

With regard to PPRL, while significant advances have been achieved in this area, there are several open research questions that need to be solved in order to make PPRL practical (Christen et al. 2014):

- So far most PPRL techniques have only investigated the linking of two databases. However, as the example scenario in Sect. 5.4.2 has shown, in many real-world applications data from more than two sources need to be linked. Our recent work in multi-party PPRL (Ranbaduge et al. 2014; Vatsalan and Christen 2014) has highlighted the significant computational challenges when aiming to link data from several sources, as even when using sophisticated blocking techniques the number of candidate record sets to be compared increases exponentially with the number of parties involved. Besides these computational challenges, possible collusion between subsets of parties needs to be considered.

- Most existing PPRL techniques only employ a simple threshold-based classifier to classify record pairs into matches or non-matches. Only group linkage (Li et al. 2011) has been considered within a PPRL framework, but none of the other advanced collective and graph-based approaches discussed in Sect. 5.2.2 have so far been investigated for their applicability in PPRL. A major challenge for classification in PPRL is the use of training data for supervised learning approaches, because such data generally require access to actual sensitive attribute values.
- How to assess linkage quality and completeness has so far not been thoroughly investigated for PPRL. This is, however, a must-solve problem as otherwise it will not be possible to evaluate the efficiency and effectiveness of PPRL techniques in real-world applications, making these techniques non-practical.
- Unlike for measuring linkage performance and quality, where standard measurements, such as run-time, reduction ratio, pairs completeness, pairs quality, precision, recall, or accuracy can be used (Christen 2012a), there are currently no standards available for measuring privacy for PPRL. Different measures have been proposed and used (Vatsalan et al. 2013b, 2014), making the comparison of techniques difficult.
- Finally, no framework has been developed that allows the experimental comparison of different PPRL techniques with regard to their scalability, linkage quality, and privacy preservation. Ideally such a framework should allow researchers to easily 'plug-in' their algorithms. Related to this issue is the lack of standard test data sets, a problem that is not just specific to PPRL but to record linkage research in general (Christen 2012a; Köpcke and Rahm 2010). A possible alternative to using real-world data sets, which are difficult to obtain due to privacy and confidentiality reasons, is to use synthetic data that are generated based on the characteristics of real data (Christen and Vatsalan 2013).

Improved collaboration between domain experts, computer scientists and statisticians who work on the algorithmic aspects of record linkage is needed to obtain the best outcomes for the field of population reconstruction. Neither research area can work in isolation. While multidisciplinary research brings its own challenges, the importance of such applied research is now increasingly being recognised by research areas that traditionally have worked in isolation (Rudin and Wagstaff 2013).

5.6 Conclusions

As our society moves into the 'Big Data' era, tremendous opportunities arise for research in the social sciences to use large-scale population-based databases collected both by commercial organisations as well as government agencies. Compared to small controlled studies based on surveys and experimental set-ups, using large databases can help overcome sampling bias and potentially reduce

costs. In an analogy to genomics and bioinformatics, Kum et al. (2013) recently proposed the notion of the 'social footprint' or 'social genome', and the field of 'population informatics' which deals with the collection, integration, and analysis of data about people gathered from many different domains, including healthcare, education, employment, finance, and so on. Reconstructing a population from such data, and enriching existing (census) data collections with such external data, will allow insights into many aspects of today's societal challenges.

National census agencies are also realising both the challenges and opportunities that matching their data with external, possibly commercial, databases can bring (Baffour et al. 2013; Office for National Statistics 2013). The acquisition of data from a variety of organisations is, however, a complicated process that involves negotiations with various partners. Privacy and confidentiality, as well as data quality issues, need to be considered carefully. As computers become more powerful, the computational challenges of linking large databases become less of an issue compared to non-technical challenges such as obtaining access to the data required for certain studies, or communication between researchers from different domains.

Nevertheless, research into techniques that allow efficient and effective population reconstruction based on data linked from a variety of sources will likely not only attract more interest from academia, but also from governments and private sector organisations. Understanding the structures and characteristics of populations, and how they change over time, becomes more valuable for organisations in an ever more competitive environment, where a better understanding of their data can give an organisation the competitive edge it needs to be successful (Siegel 2013).

Acknowledgments The authors would like to thank Mac Boot (The Australian National University) and Vassilios S. Verykios (Hellenic Open University) for their contributions to the work presented in this chapter.

References

Al-Lawati, A., Lee, D., & McDaniel, P. (2005). Blocking-aware private record linkage. In International Workshop on Information Quality in Information Systems (pp. 59–68). Baltimore.

Antonie, L., Inwood, K., Lizotte, D. J., & Ross, J. A. (2014a). Tracking people over time in 19th century Canada for longitudinal analysis. *Machine Learning, 95*, 129–146.

Antonie, L., Inwood, K., & Ross, A. (2014b). Dancing with dirty data: Problems in the extraction of life-course evidence from historical censuses. In *Population Reconstruction*.

Arasu, A., Götz, M., & Kaushik, R. (2010). On active learning of record matching packages. In ACM SIGMOD (pp. 783–794). Indianapolis.

Atallah, M. J., Kerschbaum, F., & Du, W. (2003). Secure and private sequence comparisons. In ACM Workshop on Privacy in the Electronic Society (pp. 39–44). Washington, DC.

Baffour, B., King, T., & Valente, P. (2013). The modern census: Evolution, examples and evaluation. *International Statistical Review, 81*(3), 407–425.

Bellare, K., Iyengar, S., Parameswaran, A. G., & Rastogi, V. (2012). Active sampling for entity matching. In ACM SIGKDD (pp. 1131–1139). Beijing.

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data, 1*(1), 5.

Bilenko, M., Kamath, B., & Mooney, R. J. (2006). Adaptive blocking: Learning to scale up record linkage. In IEEE ICDM (pp. 87–96). Hong Kong.

Block, W. C., & Star, D. L. (1995). Data entry and verification. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 28*(1), 63–65.

Bloothooft, G. (1995). Multi-source family reconstruction. *History and computing, 7*(2), 90–103.

Bonomi, L., Xiong, L., Chen, R., & Fung, B. (2012). Frequent grams based embedding for privacy preserving record linkage. In CIKM (pp. 1597–1601). Maui, Hawaii.

Chiang, Y. H., Doan, A., & Naughton, J. F. (2014). Tracking entities in the dynamic world: A fast algorithm for matching temporal records. *PVLDB, 7*(6).

Christen, P. (2006). A comparison of personal name matching: Techniques and practical issues. In Workshop on Mining Complex Data, held at IEEE ICDM. Hong Kong.

Christen, P. (2012a). *Data Matching—Concepts and techniques for record linkage, entity resolution, and duplicate detection. Data-centric systems and applications*. Berlin: Springer.

Christen, P. (2012b). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering, 24*(9), 1537–1555.

Christen, P. (2014). Advanced record linkage methods and privacy aspects for population reconstruction. In Population Reconstruction.

Christen, P., & Gayler, R.W. (2013). Adaptive temporal entity resolution on dynamic databases. In PAKDD (Vol. 7819, pp. 558–569). Gold Coast, Australia: Springer.

Christen, P., Gayler, R. W., & Hawking, D. (2009). Similarity-aware indexing for real-time entity resolution. In ACM CIKM (pp. 1565–1568). Hong Kong.

Christen, P., & Vatsalan, D. (2013). Flexible and extensible generation and corruption of personal data. In ACM CIKM (pp. 1165–1168). San Francisco.

Christen, P., Vatsalan, D., & Verykios, V. S. (2014). Challenges for privacy preservation in data integration. *ACM Journal Data and Information Quality, 5*(1–2), 4.

Churches, T. (2003). A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers. *BMC Med Res Methodol, 3*(1), 1.

Churches, T., Christen, P., Lim, K., & Zhu, J. X. (2002). Preparation of name and address data for record linkage using hidden Markov models. *BMC Med Inform Decis Mak, 2*, 9.

Dey, D., Mookerjee, V. S., & Liu, D. (2010). Efficient techniques for online record linkage. *IEEE Transactions on Knowledge and Data Engineering, 23*(3), 373–387.

de Vries, T., Ke, H., Chawla, S., & Christen, P. (2011). Robust record linkage blocking using suffix arrays and Bloom filters. *ACM Transactions on Knowledge Discovery from Data, 5*(2), 9.

Dong, X. L., Halevy, A., & Madhavan, J. (2005). Reference reconciliation in complex information spaces. In ACM SIGMOD (pp. 85–96). Baltimore.

Draisbach, U., Naumann, F., Szott, S., & Wonneberg, O. (2012). Adaptive windows for duplicate detection. In IEEE ICDE (pp. 1073–1083). Washington, DC.

Durham, E.A. (2012). A framework for accurate, efficient private record linkage. Ph.D. thesis, Faculty of the Graduate School of Vanderbilt University, Nashville, TN.

Durham, E. A., Xue, Y., Kantarcioglu, M., & Malin, B. (2012). Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. *Information Fusion, 13*(4), 245–259.

Dwork, C. (2006). Differential privacy. Automata, languages and programming (pp. 1–12).

Efremova, J., Ranjbar-Sahraei, B., Oliehoek, F. A., Calders, T., & Tuyls, K. (2015). A baseline method for genealogical entity resolution. In: G. Bloothooft, P. Christen, K. Mandemakers, M. Schraagen (Eds.), *Population reconstruction*. Berlin: Springer.

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering, 19*(1), 1–16.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183–1210.

Fu, Z., Boot, M., Christen, P., & Zhou, J. (2014a). Automatic record linkage of individuals and households in historical census data. *International Journal of Humanities and Arts Computing, 8*(2), 204–225.

Fu, Z., Christen, P., & Zhou, J. (2014b). A graph matching method for historical census household linkage. In PAKDD (Vol. 8443, pp. 485–496). Tainan, Taiwan: Springer.

Fu, Z., Christen, P., & Boot, M. (2011a). Automatic cleaning and linking of historical census data using household information. In Workshop on Domain Driven Data Mining, held at IEEE ICDM. Vancouver.

Fu, Z., Christen, P., & Boot, M. (2011b). A supervised learning and group linking method for historical census household linkage. In AusDM, CRPIT (Vol. 121). Ballarat, Australia.

Fu, Z., Zhou, J., Christen, P., & Boot, M. (2012) Multiple instance learning for group record linkage. In PAKDD (Vol. 7301, pp. 171–182). Kuala Lumpur, Malaysia: Springer.

Fure, E. (2000). Interactive record linkage: The cumulative construction of life courses. *Demographic Research, 3*(11), 3–11.

Glasson, E., De Klerk, N., Bass, J., Rosman, D., Palmer, L. J., & Holman, D. (2008). Cohort profile: The Western Australian family connections genealogical project. *International Journal of Epidemiology, 37*(1), 30–35.

Hernandez, M. A., & Stolfo, S. J. (1995). The merge/purge problem for large databases. In ACM SIGMOD (pp. 127–138). San Jose.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). Data quality and record linkage techniques. Berlin: Springer.

Inan, A., Kantarcioglu, M., Bertino, E., & Scannapieco, M. (2008). A hybrid approach to private record linkage. In IEEE ICDE (pp. 496–505). Cancun, Mexico.

Inan, A., Kantarcioglu, M., Ghinita, G., & Bertino, E. (2010). Private record matching using differential privacy. In EDBT (pp. 123–134). Lausanne, Switzerland.

Ioannou, E., Nejdl, W., Niederée, C., & Velegrakis, Y. (2010). On-the-fly entity-aware query processing in the presence of linkage. *VLDB Endowment, 3*(1), 429–438.

Jin, L., Li, C., & Mehrotra, S. (2003). Efficient record linkage in large data sets. In DASFAA (pp. 137–146). Tokyo.

Jonas, J., & Harper, J. (2006). *Effective counterterrorism and the limited role of predictive data mining*. Policy Analysis (584) (2006).

Kalashnikov, D. V., & Mehrotra, S. (2006). Domain-independent data cleaning via analysis of entity-relationship graph. *ACM Transactions on Database Systems, 31*(2), 716–767.

Karakasidis, A., & Verykios, V. S. (2009). Privacy preserving record linkage using phonetic codes. In Fourth Balkan Conference in Informatics, IEEE (pp. 101–106). Thessaloniki, Greece.

Karakasidis, A., & Verykios, V. S. (2010). Advances in privacy preserving record linkage. In E-activity and Innovative Technology, Advances in Applied Intelligence Technologies Book Series (pp. 22–34). IGI Global.

Karakasidis, A., & Verykios, V. S. (2012). Reference table based k-anonymous private blocking. In ACM Symposium on Applied Computing (pp. 859–864). Trento, Italy.

Karakasidis, A., Verykios, V. S., & Christen, P. (2011). Fake injection strategies for private phonetic matching. In International Workshop on Data Privacy Management. Leuven, Belgium.

Karapiperis, D., & Verykios, V. S. (2014). An LSH-based blocking approach with a homomorphic matching technique for privacy-preserving record linkage. *IEEE Transactions on Knowledge and Data Engineering*.

Kejriwal, M., & Miranker, D. P. (2013). An unsupervised algorithm for learning blocking schemes. In IEEE ICDM (pp. 340–349).

Kelman, C. W., Bass, J., & Holman, D. (2002). Research use of linked health data—A best practice protocol. *Aust NZ Journal of Public Health, 26*, 251–255.

Köpcke, H., & Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data and Knowledge Engineering, 69*(2), 197–210.

Kum, H. C., Krishnamurthy, A., Machanavajjhala, A., & Ahalt, S. (2013). Population informatics: Tapping the social genome to advance society: A vision for putting 'Big Data' to work for population informatics. *Computer, PP*(99).

Kuzu, M., Kantarcioglu, M., Inan, A., Bertino, E., Durham, E., & Malin, B. (2013). Efficient privacy-aware record integration. In EDBT (pp. 167–178). Genoa, Italy.

Lee, D., Kang, J., Mitra, P., Giles, C. L., & On, B. W. (2007). Are your citations clean? *Commununications of the ACM, 50*, 33–38.

Li, F., Chen, Y., Luo, B., Lee, D., & Liu, P. (2011). Privacy preserving group linkage. In SSDBM (Vol. 6809, pp. 432–450). Portland: Springer LNCS.

Li, P., Dong, X. L., Maurino, A., & Srivastava, D. (2011). Linking temporal records. *VLDB Endowment, 4*(11), 956–967.

Lindell, Y., & Pinkas, B. (2009). Secure multiparty computation for privacy-preserving data mining. *Journal of Privacy and Confidentiality, 1*(1), 5.

Michelson, M., & Knoblock, C. A. (2006). Learning blocking schemes for record linkage. In AAAI. Boston.

Naumann, F., & Herschel, M. (2010). An introduction to duplicate detection. *Synthesis Lectures on Data Management* (vol. 3). Morgan and Claypool Publishers.

Newcombe, H. B. (1988). *Handbook of record linkage: Methods for health and statistical studies, administration, and business*. New York: Oxford University Press Inc.

Newcombe, H. B., & Kennedy, J. M. (1962). Record linkage: making maximum use of the discriminating power of identifying information. *Communications of the ACM, 5*(11), 563–566.

Newton, G. (2013). Family reconstitution in an urban context: Some observations and methods. Technical Report, University of Cambridge, CWPESH No. 12.

Office for National Statistics. (2013). Beyond 2011 matching anonymous data. Methods and Policies Report M9.

On, B. W., Koudas, N., Lee, D., & Srivastava, D. (2007). Group linkage. In IEEE ICDE (pp. 496–505). Istanbul.

Pang, C., Gu, L., Hansen, D., & Maeder, A. (2009). Privacy-preserving fuzzy matching using a public reference table. *Intelligent Patient Management, 189*, 71–89.

Quass, D., & Starkey, P. (2003). Record linkage for genealogical databases. In ACM SIGKDD Workshop on Data Cleaning, Record Linkage and Object Consolidation (pp. 40–42). Washington DC.

Ramadan, B., Christen, P., & Liang, H. (2014). Dynamic sorted neighborhood indexing for real-time entity resolution. In ADC (Vol. 8506, pp. 1–12). Brisbane: Springer LNCS.

Ranbaduge, T., Christen, P., & Vatsalan, D. (2014). Tree based scalable indexing for multi-party privacy-preserving record linkage. In AusDM, CRPIT (Vol. 158). Brisbane, Australia.

Rastogi, V., Dalvi, N., & Garofalakis, M. (2011). *Large-scale collective entity matching. VLDB Endowment, 4*, 208–218.

Ravikumar, P., Cohen, W., & Fienberg, S. (2004). A secure protocol for computing string distance metrics. In Workshop on Privacy and Security Aspects of Data Mining held at IEEE ICDM (pp. 40–46). Brighton, UK.

Reid, A., Davies, R., & Garrett, E. (2002). Nineteenth-century scottish demography from linked censuses and civil registers: A'sets of related individuals' approach. *History and Computing, 14* (1–2), 61–86.

Rudin, C., & Wagstaff, K. L. (2013). Machine learning for science and society. *Machine Learning, 95*(1), 1–9.

Ruggles, S. (2002). Linking historical censuses: A new approach. *History and Computing, 14*(1–2), 213–224.

Scannapieco, M., Figotin, I., Bertino, E., & Elmagarmid, A. K. (2007). Privacy preserving schema and data matching. In ACM SIGMOD (pp. 653–664). Beijing.

Schneier, B. (1996). *Applied cryptography: Protocols, algorithms, and source code in C* (2nd ed.). New York: Wiley.

Schnell, R., Bachteler, T., & Reiher, J. (2009). Privacy-preserving record linkage using Bloom filters. *BioMed Central Medical Informatics and Decision Making, 9*(1), 41.

Sehili, Z., Kolb, L., Borgs, C., Schnell, R., & Rahm, E. (2015). Privacy preserving record linkage with PPJoin. In BTW Conference. Hamburg.

Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. New York: Wiley.

Su, W., Wang, J., & Lochovsky, F. H. (2009). Record matching over query results from multiple web databases. *IEEE Transactions on Knowledge and Data Engineering, 22*(4), 578–589.

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems, 10*(5), 557–570.

Talburt, J.R. (2011). Entity resolution and information quality. Morgan Kaufmann.

Toxen, B. (2014). The NSA and Snowden: Securing the all-seeing eye. *Communications of the ACM, 57*(5), 44–51.

Trepetin, S. (2008). Privacy-preserving string comparisons in record linkage systems: a review. *Information Security Journal: A Global Perspective, 17*(5), 253–266.

Vatsalan, D., & Christen, P. (2012). An iterative two-party protocol for scalable privacy-preserving record linkage. In AusDM, CRPIT (Vol. 134). Sydney, Australia.

Vatsalan, D., & Christen, P. (2014). Scalable privacy-preserving record linkage for multiple databases. In ACM CIKM. Shanghai.

Vatsalan, D., Christen, P., O'Keefe, C. M., & Verykios, V. S. (2014). An evaluation framework for privacy-preserving record linkage. *Journal of Privacy and Confidentiality, 6*(1), 3.

Vatsalan, D., Christen, P., & Verykios, V. S. (2011). An efficient two-party protocol for approximate matching in private record linkage. In AusDM, CRPIT (Vol. 121). Ballarat, Australia.

Vatsalan, D., Christen, P., & Verykios, V. S. (2013a). Efficient two-party private blocking based on sorted nearest neighborhood clustering. In ACM CIKM (pp. 1949–1958). San Francisco.

Vatsalan, D., Christen, P., & Verykios, V. S. (2013b). A taxonomy of privacy-preserving record linkage techniques. *Information Systems, 38*(6), 946–969.

Verykios, V. S., & Christen, P. (2013). Privacy-preserving record linkage. *Wiley Interdisciplinary reviews: Data Mining and Knowledge Discovery, 3*(5), 321–332.

Verykios, V. S., Karakasidis, A., & Mitrogiannis, V. K. (2009). Privacy preserving record linkage approaches. *International Journal of Data Mining, Modelling and Management, 1*(2), 206–221.

Winkler, W. E. (2006). Overview of record linkage and current research directions. Technical Report RR2006/02, US Bureau of the Census, Washington, DC.

Yakout, M., Atallah, M. J., & Elmagarmid, A. K. (2009). Efficient private record linkage. In IEEE ICDE (pp. 1283–1286). Shanghai.

Yan, S., Lee, D., Kan, M. Y., & Giles, C. L. (2007). Adaptive sorted neighborhood methods for efficient record linkage. In ACM/IEEE-CS joint conference on Digital Libraries (pp. 185–194). Vancouver.