Advanced record linkage methods and privacy aspects for population reconstruction *

Peter Christen
Research School of Computer Science
The Australian National University
Canberra ACT 0200, Australia
peter.christen@anu.edu.au

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. In this paper, we describe these major challenges of record linkage in the context of population reconstruction, outline recent developments of advanced record linkage methods, and provide directions for future research.

1 Introduction

In the past decade, record linkage has attracted much interest by researchers from various domains (Christen, 2012a; Herzog et al., 2007; Naumann and Herschel, 2010; Talburt, 2011). Also known as 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 (1969).

Computer scientists, on the other hand, 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 an author in large publication databases (Lee et al., 2007).

Social scientists working in the area of demographics and genealogy have also employed record linkage techniques, most often by using historical census data (Fure, 2000; Quass and Starkey, 2003; Ruggles, 2002). The aim of such linkages is to identify and link not just individuals across two databases, but rather to create complete family trees over significant periods of time (Antonie et al., 2013; Bloothooft, 1995; Fu et al., 2014a). Such reconstructed 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 often based on census returns, are:

*Work done in collaboration with Zhichun (Sally) Fu (ANU), Dinusha Vatsalan (ANU), Mac Boot (ANU), and Vassilios S. Verykios (Hellenic Open University). The author would like to thank Sally and Mac for their feedback, and Dinusha for her thorough proof-reading and suggestions.

- 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 sub-set 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 19th century. For each individual they usually included the name, age, gender, relationship to the head of household, and occupation. If available, 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.

- Historical census returns were 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 Section 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 will be discussed in Section 3.

In the following two sections we provide brief overviews of methods and techniques that have been developed in recent years with the aim to make record linkage applications more scalable and more accurate, and to facilitate linking records across organisations without the need to reveal private or confidential information.

2 Advanced Record Linkage Methods

In the past decade a variety of novel techniques have been developed that allow the linking of large databases. The two main areas of research have been to improve scalability and linkage quality.

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 compared with each other. The sorted neighbourhood approach (Hernandez and Stolfo, 1995) sorts a database according to a *sorting criteria* (usually a set of concatenated attribute values) 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 blocking, are all 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). 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; Ramadan et al., 2013).

Only limited experimental evaluations have been conducted to compare the performance of indexing techniques. Christen (2012b) identified that none of twelve 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 special 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 census 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; Michelson and Knoblock, 2006)).

2.2 Accurate Classification Techniques

The objective of record linkage classification is to decide if a pair or group of records are 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. The third class of *potential matches* are those pairs that require manual classification through a clerical review process.

Besides requiring an often time consuming manual clerical review step, this traditional approach has several 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 pairwise 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). 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 collective, group, and graph-based classification methods in the context of population reconstruction. Fu et al. (2011b; 2012; 2014a) have investigated group linkage methods on historical census data by treating households as groups and combining pair-wise record linkage with household linkages. 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). 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 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.

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 census 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 is increasingly considering to link contemporary databases obtained from diverse sources, privacy and confidentiality issues become crucially important. National census agencies are currently considering the use of anonymi-

sation techniques to facilitate matching their databases with records sourced from administrative data (Office for National Statistics, 2013).

While an individual 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, people are wary that their information is being collected by and shared across different organisations, especially if this is done by governments.

As an example scenario, assume a demographer who aims to investigate how mortgage stress (having to pay large sums on a regular basis to pay off 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.

While linking contemporary databases from diverse sources can allow studies at levels of detail and at scales otherwise not possible, 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. (2013b) 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 the databases at their sources and to conduct the linkage using only these encoded data (i.e. no sensitive data are ever sent to another party). 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. They can then negotiate the next steps, such as exchanging the values in certain attributes of the linked records, or, as in the above example scenario, sending selected attribute values to a third party (the researcher).

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., 2013b). The first considered the exact matching of attribute values only. 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, and then compare these 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 these first generation 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.

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., 2013b; Verykios et al., 2009). A variety of techniques have been investigated, including Bloom filters (Schnell et al., 2009; Vatsalan and Christen, 2012), phonetic encoding (Karakasidis and Verykios, 2009), random and public reference values (Karakasidis et al., 2011; Pang et al., 2009; Vatsalan et al., 2011), embedding spaces (Scannapieco et al., 2007; Yakout et al., 2009), and secure multi-party computation (Atallah et al., 2003; Inan et al., 2008; Li et al., 2011a; Ravikumar et al., 2004).

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; Inan et al., 2010; Karakasidis and Verykios, 2012; Bonomi et al., 2012; Vatsalan et al., 2013a; Durham, 2012; Kuzu et al., 2013). Different techniques have again been developed which combine traditional indexing techniques (Christen, 2012b) with encoding, perturbation, or cryptographic approaches (Vatsalan et al., 2013b). Thus far, only a few small comparative studies of such techniques have been published (Durham, 2012; Vatsalan et al., 2013a).

4 Research Directions

Most advanced record linkage techniques have been developed by computer science researchers. The focus of these techniques was not on data that contain personal information, as is generally required for population reconstruction, but often on bibliographic data. Based on this, the following research directions can be identified[1]:

- 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 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 (Christen and Gayler, 2013; Li et al., 2011b). 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), specifically in the context of population reconstruction, could lead to significant reduction in the manual efforts currently required with traditional record linkage approaches. Furthermore, how to visualise 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 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, will significantly reduce the time consuming and cumbersome process of manual data cleaning that is still commonly required today. Ideally, the requirement of training such learning algorithms is minimised by for example employing active learning or bootstrapping approaches where increasingly accurate models are trained in an iterative fashion (Churches et al., 2002). Additionally, ideally such learning techniques are transferable from one domain to another, or can be re-trained with little 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.

[1] Note these are from the viewpoint of a computer scientist who had limited exposure to the work of social scientists.

- So far most PPRL techniques have only investigated the linking of two databases. However, as the example scenario in Section 3 has shown, in many real-world applications data from more than two sources need to be linked. Besides computational challenges, possible collusion between sub-sets 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., 2011a) has been considered within a PPRL framework, but none of the other advanced collective and graph-based approaches discussed in Section 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. As a consequence, different measures have been proposed and used (Vatsalan et al., 2013b), 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 record linkage research in general (Christen, 2012a). 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 and 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 multi-disciplinary 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 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 have also started to realise both the challenges and opportunities that matching their data with external, possibly commercial, databases brings (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 ac-

cess 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 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).

References

Ali Al-Lawati, Dongwon Lee, and Patrick McDaniel. 2005. Blocking-aware private record linkage. In *International Workshop on Information Quality in Information Systems*, pages 59–68, Baltimore.

Luiza Antonie, Kris Inwood, Daniel J. Lizotte, and J. Andrew Ross. 2013. Tracking people over time in 19th century Canada for longitudinal analysis. *Machine Learning*.

Arvind Arasu, Michaela Götz, and Raghav Kaushik. 2010. On active learning of record matching packages. In *ACM SIGMOD*, pages 783–794, Indianapolis.

Mikhail J. Atallah, Florian Kerschbaum, and Wenliang Du. 2003. Secure and private sequence comparisons. In *ACM Workshop on Privacy in the Electronic Society*, pages 39–44, Washington, DC.

Bernard Baffour, Thomas King, and Paolo Valente. 2013. The modern census: Evolution, examples and evaluation. *International Statistical Review*, 81(3):407–425.

Indrajit Bhattacharya and Lise Getoor. 2007. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1(1).

Mikhail Bilenko, Beena Kamath, and Raymond J. Mooney. 2006. Adaptive blocking: Learning to scale up record linkage. In *IEEE ICDM*, pages 87–96, Hong Kong.

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

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

Luca Bonomi, Li Xiong, Rui Chen, and Benjamin Fung. 2012. Frequent grams based embedding for privacy preserving record linkage. In *CIKM*, pages 1597–1601, Maui, Hawaii.

Peter Christen and Ross W. Gayler. 2013. Adaptive temporal entity resolution on dynamic databases. In *PAKDD, Springer LNAI*, volume 7819, pages 558–569, Gold Coast, Australia.

Peter Christen and Dinusha Vatsalan. 2013. Flexible and extensible generation and corruption of personal data. In *ACM CIKM*, pages 1165–1168, San Francisco.

Peter Christen, Ross W. Gayler, and David Hawking. 2009. Similarity-aware indexing for real-time entity resolution. In *ACM CIKM*, pages 1565–1568, Hong Kong.

Peter Christen. 2012a. *Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Appl. Springer.

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

Tim Churches, Peter Christen, Kim Lim, and Justin X. Zhu. 2002. Preparation of name and address data for record linkage using hidden Markov models. *BMC Med Inform Decis Mak*, 2(9).

Tim Churches. 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).

Timothy de Vries, Hui Ke, Sanjay Chawla, and Peter Christen. 2011. Robust record linkage blocking using suffix arrays and Bloom filters. *ACM Transactions on Knowledge Discovery from Data*, 5(2).

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

Xin L. Dong, Alon Halevy, and Jayant Madhavan. 2005. Reference reconciliation in complex information spaces. In *ACM SIGMOD*, pages 85–96, Baltimore.

Uwe Draisbach, Felix Naumann, Sascha Szott, and Oliver Wonneberg. 2012. Adaptive windows for duplicate detection. In *IEEE ICDE*, pages 1073–1083, Washington, DC.

Elizabeth A. Durham, Yuan Xue, Murat Kantarcioglu, and Bradley Malin. 2012. Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. *Information Fusion*, 13(4):245–259.

Elizabeth A. Durham. 2012. *A Framework for Accurate, Efficient Private Record Linkage*. Ph.D. thesis, Faculty of the Graduate School of Vanderbilt University, Nashville, TN.

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

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

Zhichun Fu, Peter Christen, and Mac Boot. 2011a. Automatic cleaning and linking of historical census data using household information. In *Workshop on Domain Driven Data Mining, held at IEEE ICDM*, Vancouver.

Zhichun Fu, Peter Christen, and Mac Boot. 2011b. A supervised learning and group linking method for historical census household linkage. In *AusDM, CRPIT*, volume 121, Ballarat, Australia.

Zhichun Fu, Jun Zhou, Peter Christen, and Mac Boot. 2012. Multiple instance learning for group record linkage. In *PAKDD, Springer LNAI*, volume 7301, pages 171–182, Kuala Lumpur, Malaysia.

Zhichun Fu, Mac Boot, Peter Christen, and Jun Zhou. 2014a. Automatic record linkage of individuals and households in historical census data. *International Journal of Humanities and Arts Computing*.

Zhichun Fu, Peter Christen, and Jun Zhou. 2014b. A graph matching method for historical census household linkage. In *PAKDD*, Tainan, Taiwan.

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

Emma Glasson, Nick De Klerk, John Bass, Diana Rosman, Lyle J. Palmer, and D'Arcy Holman. 2008. Cohort profile: the Western Australian family connections genealogical project. *International Journal of epidemiology*, 37(1):30–35.

Mauricio A. Hernandez and Salvatore J. Stolfo. 1995. The merge/purge problem for large databases. In *ACM SIGMOD*, pages 127–138, San Jose.

Thomas N. Herzog, Fritz J. Scheuren, and William E. Winkler. 2007. *Data quality and record linkage techniques*. Springer.

Ali Inan, Murat Kantarcioglu, Elisa Bertino, and Monica Scannapieco. 2008. A hybrid approach to private record linkage. In *IEEE ICDE*, pages 496–505, Cancun, Mexico.

Ali Inan, Murat Kantarcioglu, Gabriel Ghinita, and Elisa Bertino. 2010. Private record matching using differential privacy. In *EDBT*, pages 123–134, Lausanne, Switzerland.

Ekaterini Ioannou, Wolfgang Nejdl, Claudia Niederée, and Yannis Velegrakis. 2010. On-the-fly entity-aware query processing in the presence of linkage. *VLDB Endowment*, 3(1).

Liang Jin, Chen Li, and Sharad Mehrotra. 2003. Efficient record linkage in large data sets. In *DASFAA*, pages 137–146, Tokyo.

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

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

Alexandros Karakasidis and Vassilios S. Verykios. 2009. Privacy preserving record linkage using phonetic codes. In *Fourth Balkan Conference in Informatics*, pages 101–106, Thessaloniki, Greece. IEEE.

Alexandros Karakasidis and Vassilios S. Verykios. 2010. Advances in privacy preserving record linkage. In *E-activity and Innovative Technology, Advances in Applied Intelligence Technologies Book Series*, pages 22–34. IGI Global.

Alexandros Karakasidis and Vassilios S Verykios. 2012. Reference table based k-anonymous private blocking. In *ACM Symposium on Applied Computing*, pages 859–864, Trento, Italy.

Alexandros Karakasidis, Vassilios S. Verykios, and Peter Christen. 2011. Fake injection strategies for private phonetic matching. In *International Workshop on Data Privacy Management*, Leuven, Belgium.

Chris W. Kelman, John Bass, and D'Arcy Holman. 2002. Research use of linked health data – A best practice protocol. *Aust NZ Journal of Public Health*, 26:251–255.

Hye-Chung Kum, Ashok Krishnamurthy, Ashwin Machanavajjhala, and Stanley Ahalt. 2013. Population informatics: Tapping the social genome to advance society: A vision for putting 'Big Data' to work for population informatics. *Computer*, PP(99).

Mehmet Kuzu, Murat Kantarcioglu, Ali Inan, Elisa Bertino, Elizabeth Durham, and Bradley Malin. 2013. Efficient privacy-aware record integration. In *EDBT*, pages 167–178, Genoa, Italy.

Dongwon Lee, Jaewoo Kang, Prasenjit Mitra, C. Lee Giles, and Byung-Won On. 2007. Are your citations clean? *Commununications of the ACM*, 50:33–38.

Fengjun Li, Yuxin Chen, Bo Luo, Dongwon Lee, and Peng Liu. 2011a. Privacy preserving group linkage. In *SSDBM, Springer LNCS*, volume 6809, pages 432–450, Portland.

Pei Li, Xin L. Dong, Andrea Maurino, and Divesh Srivastava. 2011b. Linking temporal records. *VLDB Endowment*, 4(11).

Matthew Michelson and Craig A. Knoblock. 2006. Learning blocking schemes for record linkage. In *AAAI*, Boston.

Felix Naumann and Melanie Herschel. 2010. *An introduction to duplicate detection*, volume 3 of *Synthesis Lectures on Data Management*. Morgan and Claypool Publishers.

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

Howard B. Newcombe. 1988. *Handbook of record linkage: methods for health and statistical studies, administration, and business*. Oxford University Press, Inc., New York, NY, USA.

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

Byung-Won On, Nick Koudas, Dongwon Lee, and Divesh Srivastava. 2007. Group linkage. In *IEEE ICDE*, pages 496–505, Istanbul.

Chaoyi Pang, Lifang Gu, David Hansen, and Anthony Maeder. 2009. Privacy-preserving fuzzy matching using a public reference table. *Intelligent Patient Management*, pages 71–89.

Dallan Quass and Paul Starkey. 2003. Record linkage for genealogical databases. In *ACM SIGKDD Workshop on Data Cleaning, Record Linkage and Object Consolidation*, pages 40–42, Washington DC.

Banda Ramadan, Peter Christen, Huizhi Liang, Ross W. Gayler, and David Hawking. 2013. Dynamic similarity-aware inverted indexing for real-time entity resolution. In *PAKDD Workshops*, Gold Coast, Australia.

Vibhor Rastogi, Nilesh Dalvi, and Minos Garofalakis. 2011. Large-scale collective entity matching. *VLDB Endowment*, 4:208–218.

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

Cynthia Rudin and Kiri L Wagstaff. 2013. Machine learning for science and society. *Machine Learning*.

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

Monica Scannapieco, Ilya Figotin, Elisa Bertino, and Ahmed K. Elmagarmid. 2007. Privacy preserving schema and data matching. In *ACM SIGMOD*, pages 653–664, Beijing.

Rainer Schnell, Tobias Bachteler, and Jörg Reiher. 2009. Privacy-preserving record linkage using Bloom filters. *BioMed Central Medical Informatics and Decision Making*, 9(1).

Eric Siegel. 2013. *Predictive Analytics: The Power to Predict who Will Click, Buy, Lie, Or Die*. John Wiley and Sons.

Weifeng Su, Jiying Wang, and Frederick H. Lochovsky. 2009. Record matching over query results from multiple web databases. *IEEE Transactions on Knowledge and Data Engineering*, 22(4):578–589.

John R. Talburt. 2011. *Entity Resolution and Information Quality*. Morgan Kaufmann.

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

Dinusha Vatsalan and Peter Christen. 2012. An iterative two-party protocol for scalable privacy-preserving record linkage. In *AusDM, CRPIT*, volume 134, Sydney, Australia.

Dinusha Vatsalan, Peter Christen, and Vassilios S. Verykios. 2011. An efficient two-party protocol for approximate matching in private record linkage. In *AusDM, CRPIT*, volume 121, Ballarat, Australia.

Dinusha Vatsalan, Peter Christen, and Vassilios S. Verykios. 2013a. Efficient two-party private blocking based on sorted nearest neighborhood clustering. In *ACM CIKM*, pages 1949–1958, San Francisco.

Dinusha Vatsalan, Peter Christen, and Vassilios S. Verykios. 2013b. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969.

Vassilios S. Verykios and Peter Christen. 2013. Privacy-preserving record linkage. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(5):321–332.

Vassilios S. Verykios, Alexandros Karakasidis, and Vassilios K. Mitrogiannis. 2009. Privacy preserving record linkage approaches. *Int. J. of Data Mining, Modelling and Management*, 1(2):206–221.

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

Mohamed Yakout, Mikhail J. Atallah, and Ahmed K. Elmagarmid. 2009. Efficient private record linkage. In *IEEE ICDE*, pages 1283–1286, Shanghai.

Su Yan, Dongwon Lee, Min-Yen Kan, and C. Lee Giles. 2007. Adaptive sorted neighborhood methods for efficient record linkage. In *ACM/IEEE-CS joint conference on Digital Libraries*, pages 185–194, Vancouver.