Li and Wang *Security Informatics* (2015) 4:6
DOI 10.1186/s13388-015-0021-0

Security Informatics
a SpringerOpen Journal

RESEARCH Open Access

A framework of identity resolution: evaluating identity attributes and matching algorithms

Jiexun Li[1] and Alan G. Wang[2*]

Abstract

Duplicate and false identity records are quite common in identity management systems due to unintentional errors or intentional deceptions. Identity resolution is to uncover identity records that are co-referent to the same real-world individual. In this paper we introduce a framework of identity resolution that covers different identity attributes and matching algorithms. Guided by social identity theories, we define three types of identity cues, namely personal identity attributes, social behavior attributes, and social relationship attributes. We also compare three matching algorithms: pair-wise comparison, transitive closure, and collective clustering. Our experiments using synthetic and real-world data demonstrate the importance of social behavior and relationship attributes for identity resolution. In particular, a collective identity resolution technique, which captures all three types of identity attributes and makes matching decisions on identities collectively, is shown to achieve the best performance among all approaches.

Keywords: Identity resolution; Social behavior; Social relationship; Collective clustering

Introduction

The world is moving away from paper-based documents to electronic records. Due to the ease of generating identity records and lack of sufficient verification or validation during data entry processes, duplicate and false identity records become quite common in electronic systems. In many practices of identity management, especially those that require integrating multiple data sources, it is often inevitable and tedious to deal with the problem of identity duplication. Particularly, finding an effective solution to this problem is extremely critical in fighting crime and terrorism to enforce national security. Criminals and terrorists often assume fake identities using either fraudulent or legitimate means so as to hide their true identity. In a number of cases documented by government reports, terrorists in different countries are known to commit identity crimes, such as falsifying passports and baptismal certificates, to facilitate their financial operations and execution of attacks, either in the real world or in the cyber space [1, 2]. The problem of an individual having multiple identities can easily mislead intelligence and law enforcement investigators [3]. Therefore, to determine whether an individual is who they claim to be is essential in the mission of counterterrorism to identify potential terrorists and prevent terrorism acts from occurring [4].

Identity resolution is a process of semantic reconciliation that determines whether a single identity is the same when being described differently [5]. The goal of resolution is to detect duplicate identity records that refer to the same individual in the real world. Over the years, researchers in the areas of database and statistics have proposed many different techniques to tackle this problem. Traditional resolution techniques rely on key attributes such as identification numbers, names and date-of-birth to detect matches. These attributes are commonly used for they are simple describers of an individual and often available in most record management systems [6, 7]. However, such personal identity attributes also vary in terms of availability and reliability across different systems. These attributes are not always accurate due to various reasons such as unintentional entry errors and intentional deception [6]. In the context of cybersecruity, it is much easier and more common for criminals to fake identities to cover their traces. In order to improve resolution accuracy, several recently proposed resolution techniques have taken into account social

* Correspondence: alanwang@vt.edu
[2]Pamplin College of Business, Virginia Tech, Blacksburg, VA 24061, USA
Full list of author information is available at the end of the article

contextual information, in addition to traditionally used descriptive identity attributes, as new evidence for identity matching [5, 8, 9]. Social contextual information, such as employment history, credit history and friendship networks, has shown to be effective in identity resolution because it is very difficult to be manipulated [10].

A review of related work leads us to believe that, although a variety of identity resolution techniques have been developed, there is not a comprehensive framework that covers various types of identity attributes and matching algorithms. This leads to the main goal of our study. In this paper, we provide an overview of various identity attributes that are useful to computational identity resolution. We develop an identity resolution framework that comprehensively considers both profile-based personal identity attributes and social identity attributes, especially with a focus on the latter. Our framework also considers different computational algorithms for matching identities, ranging from simple pairwise matching to collective clustering. Compared to existing identity resolution research, our proposed framework has the following advantages. First, the framework is generic and applicable to most record management system without introducing an overhaul in the database schema. Second, the integration of both personal and social attributes provides complementary and convincing evidence for identity resolution. Third, the collective resolution technique makes joint decisions on matching identities and therefore leads to more accurate results.

The remainder of the paper is organized as follows. We first review identity theories and existing resolution techniques. Next, we introduce our proposed identity resolution framework including several matching techniques based on various attributes and algorithms. We report a comparative study of several resolution techniques in the framework using a synthetic dataset and a real-world dataset. Finally, we conclude the paper with a summary and a discussion on future research directions.

Literature review

In this section, we review existing identity resolution approaches, with a focus on the identity attributes and matching techniques adopted.

Entity resolution and identity resolution

Identity resolution is a special type of entity resolution that specializes in identity management. Entity resolution is also known as record linkage and deduplication in the areas of statistics and database management. Record linkage, originated in the statistics community, is used to identify those records in one or multiple datasets that refer to the same real-world entity [11]. The very same task is often called and studied as record deduplication in database and artificial intelligence communities [12–14]. Given a number of records that are comprised of multiple fields, these techniques determine whether two records match by comparing the values in corresponding fields. A good survey on individual field matching models for record deduplication can be found in [15].

Entity resolution techniques can be extended to different contexts, e.g., identity resolution in identity management. However, as a special type of entity resolution, identity resolution can be very complex due to the special data characteristics of identity records. First, identity resolution, especially in the intelligence and law enforcement communities, often suffers greatly from the missing data problem [7]. Missing values, if present in many fields of a record, can present a big challenge for identity resolution techniques. Second, identity resolution needs to handle not only duplicates caused by entry errors or data ambiguities but also intentional identity fraud and deception, which tend to be hidden and concealed. Third, identity resolution techniques may need to be adjustable to different evaluation criteria. For instance, false positives may be less tolerable than false negatives for identity authentication that grants access to a critical facility. In contrast, a high false positive rate may not be a big concern when a detective searches for records related to a crime suspect with limited information. Therefore, accurate identity resolution requires a careful design that considers the special characteristics of identity records.

Identity attributes

Based on the identity theories from the social science literature, an individual's identity is considered to have two basic components, namely a personal identity and a social identity. A personal identity is one's self-perception as an individual, whereas a social identity is one's biographical history that builds up over time [16]. In particular, one's personal identity may include personal information given at birth (e.g., name, date and place of birth), personal identifiers (e.g., social security number), physical descriptions (e.g., height, weight), and biometric information (e.g., fingerprint, DNA). In contrast, a social identity is concerned with one's existence in a social context. Social identity theories consist of psychological and sociological views. The psychological view defines a social identity as one's self-perception as a member of certain social groups such as nation, culture, gender identification, and employment [17, 18]. The sociological view focuses on "the relationships between social actors who perform mutually complementary roles (e.g., employer-employee, doctor-patient)" [19]. While psychological view deals with large-scale groups, the sociological view emphasizes the role-based interpersonal relationships among people [20]. These two views combined together provide a

more complete concept for understanding a social identity at levels of social context.

Traditional identity resolution methods primarily rely on personal identity attributes such as name, gender, date of birth, and identification numbers mostly because they are commonly available as identifiers in record management systems. These attributes, however, may suffer from data quality problems such as unintentional errors [21], intentional deception [6], and missing data [7]. Biometric features such as fingerprints and DNA also belong to the category of personal attributes. Although they are considered as more reliable, they are not available or accessible due to issues such as high costs and confidentiality in most systems. A study conducted by the United Kingdom Home office [22] suggests that identity crimes usually involve the illegal use or alteration of those personal identity components. The low data quality in fact-based personal attributes can severely jeopardize the performance of identity resolution [7].

Individuals are not isolated but interconnected to each another in a society. The social context associated with an individual can be clues that reveal his or her undeniable identity. Recognizing the limitations of personal attributes, many recent studies have started exploiting social context information such as social behaviors and relationships for identity resolution. For example, Ananthakrishna et al. [23] introduced a method that eliminates duplicates in data warehouses using a dimensional hierarchy (e.g., city-state-country) over the link relations. This method scales up the ability of the matching technique by only comparing attribute values that have the same foreign key dependency. For example, the similarity of two identity records will be computed only when they both live in the same city. Kalashnikov et al. [24] incorporated co-affiliation and co-authorship relationships into an resolution model for reference disambiguation. Köpcke and Rahm [25] categorized entity resolution approaches into attribute value matchers and context matchers. While value matchers solely rely on descriptive attributes, context matchers consider information inferred from social interactions represented as linkages in a graph.

Resolution techniques

We can categorize existing resolution techniques into rule-based and machine learning approaches. There have been several rule-based identity resolution approaches based on matching rules encoded by domain experts. For instance, to integrate cross-jurisdictional criminal records, a simple rule can be: two identity records match only if their first name, last name, and date-of-birth (DOB) values are all identical [26]. Such exact-match heuristics tend to have high specificity but low sensitivity in detecting true matches, especially when data quality problems such as missing values, entry errors and deceptions are present. Hence, a good resolution technique must support partial-match as well to reduce false negatives. IBM's InfoSphere Identity Insight is a leading commercial software platform for entity resolution and analysis. It provides an identity analytics solution with a set of rules predefined by human experts as well as sophisticated algorithms. For example, given two identity records with identical dates of birth and last names, the system will resolve them into one if the matching score of their first names is above a threshold. The most critical asset and the biggest challenge of a rule-based system is the creation of the rule set. Creating a compressive rule sets can be highly time-consuming and expensive. Furthermore, rules could be domain-dependent and not portable across different contexts.

As an alternative to manual rule coding, machine learning can automatically extract patterns from annotated training data with annotated matching pairs and build resolution models for new identity records. Given a pair of identity records, distance/similarity measures are defined for different descriptive attributes and then combined into an overall score. If the overall distance (or similarity) score is below (or above) a pre-defined threshold, then the pair should be regarded as a match. Brown and Hagen [27] proposed a data association method for linking criminal records that possibly refer to the same suspect. This method compares two records and calculates a total similarity measure as a weighted-sum of the similarity measures of all corresponding feature values. Similarly, Wang et al. [6] proposed a record linkage algorithm for detecting deceptive identities by comparing four personal attributes (name, DOB, social security number, and address) and combining them into an overall similarity score. A supervised learning process determines a threshold for match decisions using a set of identity pairs labeled by an expert. However, these methods based on a limited number of descriptive attributes tend to fail if one or more of these attributes contains missing values [7]. Rather than simply labeling an identity pair as match or non-match, probabilistic methods for identity resolution root in the seminal work of [11]. By posing record linkage as a probabilistic classification problem, they propose a formal framework to predict the likelihood of matching identity pairs based on the agreement among attributes. Assuming conditional independence among features given the match class, the framework estimates the probabilistic parameters of the record linkage model in an unsupervised fashion. Built upon this work, many later studies follow and extend this framework by enriching the probabilistic model [7, 28–30]. These studies have shown that the probabilistic models achieve good performance for

identity matching. However, the parameters of the probabilistic models may not be accurately estimated in the absence of sufficient training data [31].

More machine learning based techniques have been proposed for the more general problem of entity resolution. Culotta and McCallum [32] constructed a conditional random field model (CRF) for record deduplication that captures inter-dependencies between different types of entities. This method, however, fails to model the explicit links among the same type of entities. Pasula et al. [9] proposed a citation matching approach based on probabilistic relational model (PRM), which is built upon dependences among entities in a relational database schema through foreign key relationships. Li et al. [3] developed a systematic approach to deriving social behavior and social relationship features for identity matching based on a database schema and PRMs. This approach can be used as a plug-and-play solution for most relational database-based record management systems. For less structured collections of record data (e.g., criminal reports, news articles), a lot of efforts for information extraction and transformation will be needed. Bhattacharya and Getoor [33] proposed a graph-based method for entity resolution. It defines a similarity measure that combines corresponding attribute similarities with graph-based relational similarity between each entity reference pair. Furthermore, Bhattacharya and Getoor [8] extended their relational resolution approach and introduced a collective entity resolution algorithm. Rather than simply making pair-wise entity comparisons, their method can derive new social information from decisions already made and incorporate it into further resolution process iteratively. More recently, in order to tackle the problem of one individual having several profiles on different social media platforms (e.g., Facebook and Twitter), researchers have developed techniques for matching user profiles. Bartunov et al. [34] developed a CRF-based approach that combines two user graphs using both user profile attributes and social linkages. All these studies have demonstrated that social information, when incorporated in matching algorithms, can improve the performance for identity resolution.

Research gap

According to our literature review, many researchers have recognized the limitations of traditional identity resolution techniques that solely rely on personal identity attributes such as name and DOB. Although these profile-based identity attributes are available in most record management systems, they are subject to data entry errors, deception, and fraud [6]. Hence, techniques based on such attributes would fail drastically in recognizing the unconventional truth like the synonym between *Osama bin Laden* and *The Prince* [35]. Several recent studies have demonstrated the benefits of utilizing social contextual information in identity resolution [3, 36]. However, most studies lack the guidance of identity theories to construct and examine different types of social attributes for identity resolution. Identity theories suggest that personal and social identity may complement each other for the purpose of identity resolution. The social aspect of identity reflects one's psychological and sociological perception of his/her own identity. Social identity attributes might be more reliable than personal profiles in that they cannot be easily altered or falsified by an individual. Furthermore, existing resolution techniques mainly employ pair-wise comparison [6, 27, 37, 38] when finding matching identity records. Entity resolution studies have shown that resolution accuracy can improve significantly if matching of related identity references are performed in a collective fashion [8]. The effectiveness of a collective approach in the context of identity resolution is yet to be examined. Built upon our previous work in [39], this study is to develop a comprehensive identity resolution framework by examining a variety of identity attributes and evaluating different matching strategies.

A framework of identity resolution

In this study, we aim to contribute to the field of identity resolution from the following three aspects:

- Developing a comprehensive framework of identity resolution that covers both the usage of identity attributes and matching algorithms;
- Examining the predictive powers of personal identity attributes and social identity features for resolution; and
- Evaluating different matching algorithms for identity resolution.

Problem definition

We define the identity resolution problem as follows. Given a set of identity references, $R = \{r_i\}$, where each reference r is described by a set of attributes $r.A_1$, $r.A_2$, …, $r.A_l$. These references correspond to a set of unknown individual $E = \{e_i\}$. Due to certain reasons (e.g., entry errors, duplicates or deceptions), multiple references may be co-referent to the same underlying individual e. We use $r.E$ to refer to the individual to which reference r refers to. Each reference r may have participated in some incident(s) (e.g., a financial transaction, a criminal misconduct, a terrorist attack) and each incident may involve one or multiple references. We use a set of hyper-edges $H = \{h_i\}$ to represent all incidents. Each incident is a hyper-edge that connects multiple references. Each hyper-edge h can also be described by a set of attributes $h.B_1$, $h.B_2$, …,

$h.B_m$. We use $h.R$ to denote the set of references involved in h. We use $r.H$ to denote the set of hyper-edges in which r participates. In addition, the participation of r in h can be described by a set of attributes $r.h.C_1$, $r.h.C_2$, …, $r.h.C_n$. The objective is to uncover the hidden set of individuals $E = \{e_i\}$ and the entity labels $r.E$ for all references given the observations of the references and their involved hyper-edges.

Figure 1 illustrates the problem of identity resolution in a graph of five reference nodes connected by three hyper-edges. In this graph, two nodes (references) r_1 and r_2 are in fact co-referent to the same person; r_4 and r_5 are co-referent to another person. Our goal is to uncover the underlying matches from the graph. In addition to simply comparing the node-level attributes of references, we should also consider the attributes related to their involved incidents, represented by hyper-edges. For example, if h_1 and h_2 share some similar patterns (e.g., incident type, time of day), or if r_1 and r_2 have similar patterns (e.g., role, activeness) in their participation in h_1 and h_2 respectively, these should be considered as supporting evidence for r_1 and r_2 being co-referent to the same individual. Furthermore, Moreover, the common neighbor r_3 shared by r_1 and r_2 can also indicate potential linkage between the pair. Furthermore, once r_1 and r_2 are determined to be co-referent, the joining of these two can bring in new contextual information for inferring the matching relationship between other identity references (e.g., r_4 and r_5). We will discuss how to mathematically represent and compute such evidence in the rest of Section "A framework of identity resolution".

To solve this problem, we develop a framework of identity resolution by considering two aspects: identity attributes and matching algorithms. Identity attributes are features that can be used to describe certain distinctive characteristics of identity references. A matching algorithm defines the computational procedure of recovering the matching entities from the given set of references. In this section we first define various types of identity attributes that can be used in identity resolution. We also discuss how we calculate the similarity for each type of identity attributes. Lastly, we introduce several matching algorithms that match identity references based on attribute similarities.

Identity attributes

To determine whether two references are co-referent, we need to compare their common attributes and calculate corresponding similarity scores for these attributes. Given our problem definition, we divide all identity attributes into three categories and describe their similarity measures as follows.

Personal identity attributes

Personal identity attributes are identifiers that are commonly used in record management systems to distinguish one person from others. In our problem definition, each reference r is represented by a node in the graph. Personal identity attributes are the node-level attributes, $r.A_1$, $r.A_2$, …, $r.A_l$, Examples

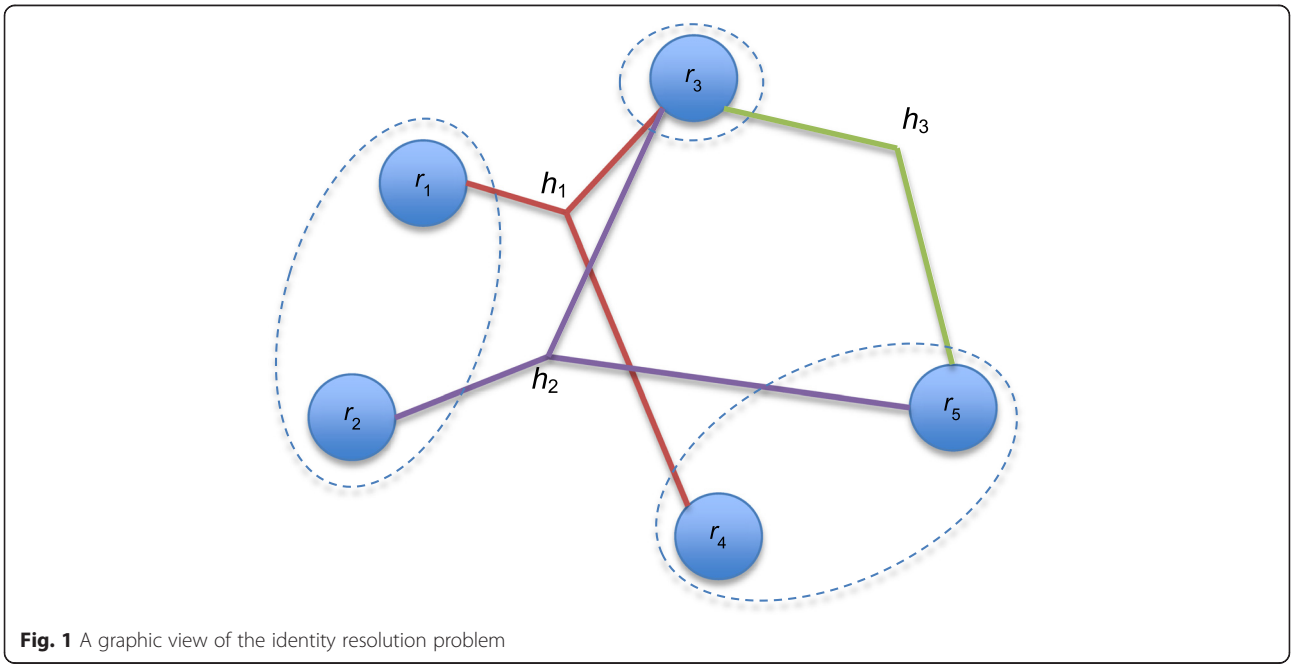

Fig. 1 A graphic view of the identity resolution problem

include but not limited to name, date of birth, social security number, and passport number. For an attribute in text format, the similarity between two attribute values can be calculated by edit distance measures such as the Levenstein Distance [40]. For an attribute in number or date format, absolute difference or percentage difference can be used as similarity measure. We use p to refer to resolution approaches based on personal identity attributes alone:

$$sim_p(r_i, r_j) = \frac{1}{l}\sum_{k=1}^{l} sim(r_i.A_k, r_j.A_k).$$

If the similarity score $sim_p(r_i, r_j)$ is above a threshold, references r_i and r_j are considered to be co-referent. These approaches tend to perform well for references with duplicates and typos, but are likely to have low recall for cases involving intentional deceptions.

Social behavior attributes

Based on social identity theories, we consider an individual's behaviors in the society as reflections of his/her underlying identity. Social behavior attributes represent the common characteristics of the social group(s) that one belongs to. They reflect the psychological view of personal identity. Most identity management systems contains information about incidents (e.g., credit card transactions, crime misconducts) related to each individuals. In our problem definition, each incident is denoted by a hyper-edge in the graph. Therefore, we use incident-based behavioral patterns to describe the characteristics of an individual. Each hyper-edge h is associated with a set of attributes $h.B_1$, $h.B_2$, ..., $h.B_m$ (e.g., transaction day/time/amount, crime type). In addition, the participation of r in h has a set of attributes $r.h.C_1$, $r.h.C_2$, ..., $r.h.C_n$ (e.g., role, start/end time). Such attributes, capturing information about how individual behave in certain incidents, should also be considered as social behavior attributes for identity resolution. It is worth noting that this type of participating attributes was not considered in previous work [8, 39], adding these attributes has enriched the social behavior attributes and made this a more comprehensive framework of identity resolution. We denote the set of hyper-edges involving reference r as $r.H$. Then, each hyper-edge $r_i.h$ involving r_i is compared with each hyper-edge $r_j.h'$ of r_j based on attributes $\{B_k\}$ and $\{C_k\}$. The overall behavior similarity between r_i and r_j, $sim_b(r_i, r_j)$, is the average of similarity scores between each hyper-edge pairs:

$$sim_b(r_i, r_j) = sim_b(r_i.H, r_j.H) = \frac{1}{|r_i.H| \times |r_j.H|} \sum_{h \in r_i.H, h' \in r_j.H} sim_b(r_i.h, r_j.h'),$$

where $|r.H|$ denotes the number of hyper-edges involving r, and $sim_b(r_i.h, r_j.h')$ is defined as the average similarity score for each attribute of $h.B_k$ and $r.h.C_k$:

$$sim_b(r_i.h, r_j.h') = \frac{1}{2mn}\left(n\sum_{k=1}^{m} sim(h.B_k, h'.B_k) + m\sum_{k=1}^{n} sim(r_i.h.C_k, r_j.h'.C_k)\right).$$

Social relationship attributes

The sociological view is the other level of a social identity that concerns with social relationships of an individual. We can define social relationship attributes to capture this aspect of social identities. If two references r_i and r_j are both related to the same reference r_k (e.g., r_i and r_j co-occur with r_k in different hyper-edges), this can be regarded as evidence that these two references are co-referent. We denote the neighborhood of a reference r as $Nbr(r)$. Then, the neighborhood similarity between two references r_i and r_j, $sim_n(r_i, r_j)$, can be defined as:

$$\begin{aligned} sim_n(r_i, r_j) &= sim_n(Nbr(r_i), Nbr(r_j)) \\ &= sim_n(r_i.H, r_j.H). \end{aligned}$$

To compute the neighborhood similarity between references r_i and r_j, we define hyper-edge neighborhood similarity $sim_n(h_i, h_j)$ between two hyper-edges h_i and h_j as a pair-wise match between their references.

Here, the relational similarity between two hyper-edges h_i and h_j is computed as the maximum of the similarity score between a reference $r \in h_i.R$ and a reference $r' \in h_j.R$:

$$sim_n(h_i, h_j) = \max_{r \in h_i.R, r' \in h_j.R}(sim_p(r, r')).$$

Furthermore, the neighborhood similarity between two references r_i and r_j is defined as:

$$sim_n(r_i, r_j) = \frac{1}{|r_i.H| \times |r_j.H|} \sum_{h \in r_i.H, h' \in r_j.H} sim_n(r_i.h, r_j.h'),$$

where $|r.H|$ still denotes the number of hyper-edges involving r. It is worth noting that the neighborhood of reference r contains r. Thus, if two references r_i and r_j do not share any common neighbor, their neighborhood similarity $sim_n(r_i, r_j)$ is equal to their personal identity attribute similarity $sim_p(r_i, r_j)$. In other words, two references having the same/similar neighbors can be regarded

as evidence to support that they are more likely to be co-referent, whereas two references not sharing a common neighbor will not be treated as evidence of them not being co-referent.

Besides, we can define a negative constraint based on social relationship, i.e., two references that co-occur in the same hyper-edge cannot refer to the same individual. It is unlikely that one real-world individual can assume two different identities in one unique incident.

Aggregated similarity score

For a pair of references r_i and r_j, we have defined three scores that measure their similarity based on personal identity, social behavior and social relationship attributes. In order to take into account different attributes for identity resolution, we need to aggregate these similarity scores to an overall score.

Here we use a weighted average method to computing the overall similarity. It is worth noting that, before being combined, similarity scores should be normalized to the same scale, e.g., in the range of [0, 1]. Thus, the overall similarity score between two references r_i and r_j is a weighted average of all similarity scores: i.e., personal identity similarity, social behavior similarity, and social neighborhood similarity:

$$AVG(r_i, r_j) = \alpha \times sim_p(r_i, r_j) + \beta \times sim_b(r_i, r_j) + (1-\alpha-\beta) \times sim_n(r_i, r_j)$$

where weights $0 \le \alpha,\ \beta \le \alpha + \beta \le 1$ and they can be adjusted to control the importance of the three similarity scores.

If $\alpha = 1$, then only personal identity similarity is considered and $AVG(r_i, r_j)$ is simply $sim_a(r_i, r_j)$. We represent this as $sim_A(r_i, r_j)$.

If $\alpha, \beta > 0$ and $\alpha + \beta = 1$, then $AVG\ (r_i, r_j)$ is the average of personal identity similarity and social behavior similarity whereas social neighborhood similarity is not considered. We represent this as $sim_B(r_i, r_j)$.

If $0 < \alpha,\ \beta \le \alpha + \beta < 1$, then all three similarity scores are considered in the average score. We represent this as $sim_R(r_i, r_j)$.

Matching algorithms

Given a collection of references and similarity measures defined above, we still design an algorithmic process to traverse and compare all identity pairs and eventually reveal all underlying individuals. Here, we describe three matching algorithms in existing resolution techniques, namely pairwise comparison, transitive closure, and collective clustering.

Pairwise comparison

Pair-wise comparison is a basic and simple procedure for entity resolution. For each pair of references r_i and r_j, we can compute the similarity score using one of the aforementioned functions. If the similarity score $sim(r_i, r_j)$ is greater than a predefined threshold θ, we conclude that r_i and r_j are co-referent. Figure 2 shows the pseudo-code of this algorithm:

Transitive closure

The outcome of pairwise comparison only tells whether each reference pair matches or not. Technically, this still has not yet uncovered the underlying individuals in the reference collection. Consider a simple example where references r_i and r_j are a match, r_j and r_k are also a match, while r_i and r_k are determined to be a non-match. In this case, the pair-wise comparison may produce conflicting resolution outcomes. Hence, to uncover the underlying individuals, an extra step is to use transitive closure and merge all related matching results. In the previous example, even though r_i and r_k are not sufficiently similar, we still consider them as a co-referent pair because r_i matches r_j and r_j matches r_k. This process should be performed iteratively until all transitive closures are reached. Finally, each transitive closure is considered as a distinct individual. Given the output from pairwise comparison, computing transitive closures is straightforward and efficient. However, it is worth noting that such a merging process lowers the threshold of the matching criteria. As a result, it tends to reduce false negatives but introduce more false positives. Figure 3 shows the pseudo-code of this algorithm. Specifically, at Step 4, the results of pairwise comparison in 3.3.1 is used to judge if two references r and r' match. This process does require recalculating any similarity scores. The minimum distance between references from two clusters is simply considered as the distance between the two clusters. In this sense, this algorithm is a single-linkage or nearest neighbor clustering method.

Collective clustering

Collective clustering is a different matching technique for identity resolution [8]. It adopts a greedy agglomerative clustering algorithm to find the most similar references (or clusters) and merge them. There are some

```
1.  For each pair of two different references (r_i, r_j) in R×R
2.      Compute their similarity sim(r_i, r_j)
3.      If sim(r_i, r_j) is greater than threshold
4.          Predict the pair (r_i, r_j) as a match
```

Fig. 2 Pseudo-code of pairwise comparison

```
1.  Initialize each reference as a cluster
2.  For each cluster c_i
3.    For each cluster c_j
4.      If ∃ r∈c_i and r'∈c_j s.t. r matches r'
5.        Merge c_i and c_j
```

Fig. 3 Pseudo-code of transitive closure

fundamental differences between this algorithm with the two introduced above.

Collective clustering does requires defining a similarity function on clusters of references. Each cluster formed is considered the same real-world individual. Unlike transitive closure that uses single-linkage clustering, collective clustering uses an average linkage approach instead. It defines the similarity between two clusters c_i and c_j as the average similarity between each reference in c_i and each reference in c_j:

$$sim(c_i,c_j) = \frac{1}{|c_i.R| \times |c_j.R|} \sum_{r\in c_j.R, r'\in c_j.R} sim(c_j.r, c_j.r'),$$

where $|c.R|$ represents the number of references in cluster c.

Figure 4 shows the pseudo-code of this algorithm. Like transitive closure, collective clustering begins with assigning each reference to a different cluster. Iteratively, this algorithm computes the between-cluster similarity and merges the most similar pair of clusters until the similarity drops below the predefined threshold. As clusters merge, one cluster can contain multiple references that are determined to be co-referent. A critical step in this algorithm is that, every time two clusters are merged, all similarity scores that involve references from the merged clusters must be recomputed. Specifically, the merging of two clusters c_i and c_j into one, which could change the linkage structures in the graph, should be regarded as new evidence for computing the similarity between other references that co-occur with those in c_i and c_j. Such an iterative computational procedure, which takes into account resolution decisions on all nodes in the graph jointly, makes this clustering algorithm collective and distinguishes it with the other pairwise comparison based algorithms.

```
1.  Initialize each reference as a cluster
2.  Compute the similarity between each cluster pair
3.  Find the cluster pair with the maximal sim*(c_i, c_j)
4.  If sim*(c_i, c_j) greater than threshold
5.    Merge c_i and c_j
6.    Go to Step 2
```

Fig. 4 Pseudo-code of collective clustering

Complexity analysis

For a collection of n references, a complete pairwise comparison approach considers all possible reference pairs as potential candidates for merging. Such a process has a complexity of $O(n^2)$. The collective clustering algorithm involves an iterative procedure of recalculating cluster similarities and therefore requires even more computation. Apart from the scaling issue, in reality most pairs will be rejected while often times only less than 1 % of pairs are real matches. Hence, certain blocking steps should be incorporated and performed before matching so as to screen out highly unlikely candidate pairs for matching. There are various blocking techniques but they often employ one simple and computationally cheap function to group references into a number of buckets. Buckets can overlap so one reference can belong to multiple buckets. Only reference pairs within the same bucket should be considered candidate pairs for comparing and merging, whereas a pair of references from two different buckets is not considered as candidate for further comparison. For collective clustering, two clusters must have all of their references belonging to the same bucket to make them a candidate pair to be compared and possibly merged. Furthermore, in the agglomerative clustering process, it is not necessary to recalculate the similarity scores for every cluster pair. Only those involving references from the merged clusters need to be recomputed. We have implemented these aforementioned strategies for reducing complexity in our evaluation. More strategies for reducing the complexity of collective clustering can be found in [8].

Comparative evaluation

In this study, we conducted a comparative evaluation of identity resolution techniques based on different identity attributes and matching algorithms. We describe our experiments and discuss our findings in this section.

Datasets

Our evaluation used both a computer-generated synthetic dataset and a real-world dataset. The synthetic dataset was used to conduct a more comprehensive empirical evaluation on the proposed framework. The real-world dataset was used to test the usefulness and reliability of the proposed framework in a real setting.

A synthetic dataset

Synthetic data allows us to embed ground truth in the data generated for algorithm testing. Here, we adopted a two-staged data generation method, as described in Fig. 5.

In the creation stage the algorithm first creates N entities with a node attribute x, a hyper-edge attribute y, and a participating attribute z. We differentiate these

Creation Stage
1. Repeat N times
2. Create entity e
3. Create attribute x to represent e's node attribute
4. Create attribute y to represent e's hyper-edge attribute
5. Create attribute z to represent e's participating attribute
6. Repeat M times
7. Choose entities e_i and e_j with $P_b(e_i, e_j)$
8. Set $e_i = Nbr(e_j)$ and $e_j = Nbr(e_i)$

Generation Stage
1. Repeat L times
2. Randomly choose entity e
3. With P_a, select reference r for e or generate r using $N(e.x, 1)$
4. Initialize hyper-edge h = <r>
5. Repeat with probability P_c
6. Randomly choose e_j from $Nbr(e)$ without replacement
7. With P_a, select reference r_j for e_j or generate r_j using $N(e_j.x,1)$
8. Add r_j to hyper-edge h
9. Assign hyper-edge h's attribute: $h.y$ = average($h.r.e.y$)
10. For each r in h
11. Assign r's attribute in h: $r.h.z = (r.e.z+h.y)/2$

Fig. 5 Pseudo-code of synthetic data generation

three attributes because personal identity attributes (x) can be easily modified or falsified whereas social behavior attributes (y and z) tend to be more consistent and less likely to change significantly over time. Next, we create M linkages among these N entities to represent their underlying social relationships. When we create these links, two entities with similar social behavior attribute y, are more likely to be connected to each other with a probability of $P_b(e_i, e_j) = P_b^{|e_i.y-e_j.y|}$, i.e., two individuals with similar interests or behavioral patterns are more likely to be related. In the generation stage, we created R hyper-edges that included a set of related entities. An entity may join a hyper-edge of its neighbor with a probability of P_c. Whenever an entity is to join a hyper-edge, we either choose an existing reference of this entity with a probability P_a or create a new reference with its attribute x value following a Gaussian distribution of $N(e.x, 1)$. Each reference r joins at least one hyper-edge. Each hyper-edge is also assigned an attribute y, which is equal to the average of all participating entities' behavior attribute y. This is to assure the hyper-edge's attribute value not quite distant from the participating entities'. Similarly, for each participating reference r in h, we assign its participating attribute as the average of the underlying entity's attribute $r.e.z$ and the hyper-edge's attribute $h.y$.

It is worth noting that this synthetic data generator is based on a method used in [8] but has two major differences. First, for each entity, we define a separate attribute y to represent one's behavior characteristics that can be reflected and observed in its involved hyper-edges. Second, our synthetic data allow a reference to join more than one hyper-edge, while in [8] each reference only joins a single hyper-edge. This is more realistic in real-world identity management scenarios and requires more cost in computing similarities between references.

In our experiments, we chose the following parameter values for generating the synthetic data: $N = 50$, $M = 100$, $L = 200$, $P_a = 0.8$, $P_b = 0.9$, $P_c = 0.6$, and the value ranges of attributes x, y and z were set to [0, 100). With such parameter settings, we generated 50 sets of synthetic data and compared different resolution approaches. Since we limit the attribute values to be a number between 0 and 100 instead of actual identity attribute values (e.g., DOB, name), we simplified the similarity calculation introduced in Section "Identity attributes" as follows:

The similarity between two hyper-edges is defined as:

$$sim(h_i, h_j) = \frac{|h_i.y - h_j.y|}{range(y)}$$

The similarity between two participations of references in hyper-edges is defined as:

$$sim(r_i.h, r_j.h') = \frac{|r_i.h.z - r_j.h'.z|}{range(z)}$$

A real dataset

Ideally, to empirically evaluate and compare these identity resolution techniques, we need a real-world dataset in which duplicate identity records exist. In previous literature of entity resolution, many researchers have used citation data as test bed [8]. In order to better differentiate the three types of identity attributes, in this study we chose to use a dataset manually extracted from a set of public web pages and news stories related to terrorism and subjectively linked entities mentioned in the articles. These pages and articles were mostly hosted by web sites of governments and news organizations. This data set was originally collected and annotated by Kubica et al. [41]. It has been used in several studies for testing link analysis and alias detection [35, 42]. In total, this dataset consists of 4088 entity names. Among them, there are 164 names of terrorists. These 164 names have been manually reviewed and recognized as co-references of 20 unique individuals including well-known names such as *Osama Bin Laden* and *Abdel Muaz*. For instance, there are 12 different names and references for *Osama Bin Laden*. In addition, this dataset also consists of 5581 links among these 4088 names, representing their co-occurrence relationships in articles.

The terrorist names are only a small subset of the entire set of 4088 names. For each terrorist name, we regard all its associated non-terrorist names as its social behavior attributes and all its linked terrorist names as its social relationship attributes.

Experimental design

n the experiments, we compared different identity resolution techniques in our framework. For identity attributes, we considered the following three sets, which cumulatively added one type of attributes on top of existing ones:

- personal identity attributes alone;
- personal identity + social behavior attributes; and
- personal identity + social behavior + social relationship attributes

For each attribute set, we conducted identity resolution using one of the three matching algorithms: pairwise comparison, transitive closure, and collective clustering. Since the collective clustering algorithm involves social linkages between individuals, our implementation of this algorithm captures all three types of attributes. Table 1 lists seven combinations of identity attributes and matching algorithms that correspond to seven identity resolution approaches we tested in our comparative experiments. Each technique is given an acronym as shown.

We evaluated the performance of each identity resolution approach by checking the correctness of the matching decisions for each reference pair. We followed most identity resolution studies and chose precision, recall and F-measure as our evaluation metrics [8]. As illustrated in Table 2, each matching decision on a pair of references is either a "match" (positive) or a "non-match" (negative). A decision can be either true or false.

Based on the decision outcomes, precision, recall, and *F*-measure are defined as follows:

Table 1 Experiment design: identity attributes vs. matching algorithms

	Attributes		
Algorithms	Personal Identity	Personal Identity + Social Behavior	Personal Identity + Social Behavior + Social Relationship
Pair-wise Comparison	A	B	R
Transitive Closure	A*	B*	R*
Collective Resolution	–	–	CR

Table 2 Possible outcomes of matching decisions

	Decision	
Reality	Match	Non-match
Match	True positive (*TP*)	False negative (*FN*)
Non-match	False positive (*FP*)	True negative (*TN*)

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F\text{–}measure = \frac{2 \times precision \times recall}{precision + recall}$$

Results and discussion

Results for synthetic data

In our experiment on the synthetic dataset, we used the weighted average approach to aggregate all similarity scores. We experimented with different parameter settings so as to have a more comprehensive comparison of different identity resolution techniques. We also experimented with different weight parameter values for α ($0.0 \sim 1.0$) and β ($0.0 \sim 0.5$). Table 3 summarizes the best performance for each of the seven resolution techniques. Solely using personal identity attributes (represented by x in our synthetic data), method A tends to have high recall (92.62 %) but very low precision (9.43 %). If two identity references have very similar personal identity attributes, A tends to consider them co-referent. Method B takes into account social behavior attributes represented by y and z in our synthetic data. It achieved significantly higher precision (85.66 %) at the cost of lower recall (37.71 %) but higher F-measure (44.37 %) than A. Furthermore, method R added social relationship attributes and achieved better performance than B in terms of recall (37.87 %) and F-measure (50.41 %). The three methods using transitive closures (A*, B*, R*) did not seem to give significantly better results than their corresponding pairwise comparison methods. Notably, the collective resolution algorithm CR achieved the highest

Table 3 Peformance of identity resolution for synthetic data

Method	α	β	1-α-β	Precision	Recall	F-measure
A	1.0	0.0	0.0	9.43 %	92.62 %	16.97 %
A*	1.0	0.0	0.0	7.72 %	94.27 %	14.16 %
B	0.7	0.3	0.0	85.66 %	30.71 %	44.37 %
B*	0.7	0.3	0.0	85.02 %	31.21 %	44.71 %
R	0.5	0.25	0.25	79.22 %	37.87 %	50.41 %
R*	0.5	0.25	0.25	77.11 %	38.64 %	50.44 %
CR	0.5	0.25	0.25	87.95 %	44.15 %	58.17 %

precision (87.95 %) and the highest F-measure (58.17 %), which were both significantly higher than those of the other six methods. Therefore, our experiments on the synthetic data demonstrated that, by considering all three types of identity attributes and matching references in a collective and iterative manner, CR is a more effective method for identity resolution.

Results for real-world data

In the empirical evaluation on the real-world dataset, we also experimented with aggregated similarity scores under different parameter settings so as to have a more comprehensive comparison of the identity resolution techniques. Specifically, we compared different weight values for α ($0.0 \sim 1.0$) and β ($0.0 \sim 0.5$). Table 4 shows the best performance of the identity resolution techniques achieved and their corresponding parameter settings.

Table 4 summaries the performance of identity resolution methods using weighted average scores. When $\alpha = 0.6$, $\beta = 0.0$ and $1\text{-}\alpha\text{-}\beta = 0.4$, the collective clustering resolution (CR) achieved the highest precision (99.34 %), recall (54.30 %), and F-measures (70.22 %).

Discussion

Our experiment results for the synthetic data show that, with the same matching algorithm, using personal identity attributes alone may achieve higher recall but at the cost of low precision and F-measure. As social behavior attributes being taken into account, we observed a significant boost in both precision and F-measure. This shows that social behavior attributes do contribute to the performance improvement of identity resolution by reducing false positives. Furthermore, when social relationship attributes were also considered, there were some minor drops in precision but with generally higher *F*-measure. As confirmed by results for both synthetic and real data, the approaches using all three types of identity attributes (*R*, *R** and *CR*) outperformed its baselines in terms of F-measure. In particular, the collective clustering algorithm (*CR*), which not only considers all types of personal and social attributes but makes matching decisions on identities in a collective fashion, was demonstrated to be the optimal methods that outperform all others.

Table 4 Peformance of identity resolution for real-world data

Method	α	β	1-α-β	Precision	Recall	F-measure
A	1.0	0.0	0.0	48.59 %	18.50 %	26.79 %
A*	1.0	0.0	0.0	48.59 %	18.50 %	26.79 %
B	0.9	0.1	0.0	57.98 %	16.47 %	25.65 %
B*	0.9	0.1	0.0	57.98 %	16.47 %	25.65 %
R	0.1	0.0	0.9	97.29 %	38.54 %	55.21 %
R*	0.1	0.0	0.9	97.29 %	38.54 %	55.21 %
CR	0.6	0.0	0.4	99.34 %	54.30 %	70.22 %

Notably, the optimal parameter values (i.e., α and β) varied in our experiments across the two datasets and different algorithms. In general, α was greater than β, indicating personal identity attributes being more important than social relationship attributes. Particularly for the real-world dataset, β values were very small for algorithms B and B* and zero for R, R*, and CR. The results show that social behavior attributes did not help much in finding matching terrorists. In our study, we considered a terrorist's associated non-terrorist names as his/her social behavior attributes. The low β values may be caused by the nature of terrorist incidents, which each time involve different non-terrorist people such as victims. The weight for social relationship attributes, 1-α -β, was high in R, R*, and CR, which showed the important of social relationship attributes in finding matching terrorists.

Conclusions and future directions

In this paper we introduced a framework of identity resolution techniques that utilizes different identity attributes and matching algorithms. Guided by existing identity theories, we defined and examined three types of identity cues, namely personal identity attributes, social behavior attributes, and social relationship attributes, for identity resolution. We also evaluated three matching algorithms: pair-wise comparison, transitive closure, and collective clustering. Our experimental results showed that both social behavior and relationship attributes improved the performance of identity resolution as compared to using personal identity attributes alone. In particular, the collective identity resolution algorithm, which considers all three types of identity attributes in a collective clustering process, was shown to achieve the best performance among all approaches.

Since this study mainly focuses on comparing the three types of identity attributes with three identity resolution algorithms. These three algorithms are all similarity-based unsupervised learning methods. In our future research, we plan to compare them with other existing identity resolution algorithms (e.g., supervised learning algorithms). Another limitation of this study is the type and size of datasets employed in experiments. To validate, improve, and operationalize these identity resolution techniques, we plan to test them on real-world identity datasets with larger scales and more attributes. Furthermore, another direction of our future research is to further optimize the complexity of identity matching algorithms for more efficient processing.

Abbreviations
CRF: Conditional random field model; PRM: Probabilistic relational model; CR: Collective resolution.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JL conceived of the study, carried out the research design, conducted the experiments and drafted the manuscript. AGW co-designed the proposed approach, participated in the experimental study and drafted the manuscript. Both authors read and approved the final manuscript.

Authors' information
JL is an Assistant Professor in Business Information Systems, College of Business, at Oregon State University. He earned his Ph.D. in Management Information Systems from the University of Arizona. His research interests include data mining, business analytics, social media analytics, and health informatics. He has published in Journal of Management Information Systems, Decision Support Systems, IEEE Transactions, Journal of the American Society for Information Science and Technology, Journal of the Association for Information Systems, Bioinformatics, Communications of the ACM, Expert Systems with Applications, and Information Systems Frontiers. GAW is an Associate Professor in the Department of Business Information Technology, Pamplin College of Business, at Virginia Tech. He received his Ph.D. in Management Information Systems from the University of Arizona. His research interests include heterogeneous data management, data cleansing, data mining and knowledge discovery, and decision support systems. He has published in Production and Operations Management, Communications of the ACM, IEEE Transactions of Systems, Man and Cybernetics (Part A), IEEE Computer, Group Decision and Negotiation, and Journal of the American Society for Information Science and Technology. Dr. Wang is a member of the Association for Information Systems (AIS), the Decision Sciences Institute (DSI), and the Institute of Electrical and Electronics Engineers (IEEE).

Author details
[1]College of Business, Oregon State University, Corvallis, OR 97331, USA.
[2]Pamplin College of Business, Virginia Tech, Blacksburg, VA 24061, USA.

Received: 23 April 2015 Accepted: 13 July 2015
Published online: 28 July 2015

References

1. TH Kean, CA Kojm, P Zelikow, JR Thompson, S Gorton, TJ Roemer, JS Gorelick, JF Lehman, FF Fielding, B Kerrey, The 9/11 Commission Report. 2004. URL: http://govinfo.library.unt.edu/911/report/index.htm
2. U.S. Department of State: Country Reports on Terrorism 2006. 2007. URL: http://www.state.gov/j/ct/rls/crt/2006/
3. J Li, GA Wang, H Chen, Identity matching using personal and social identity features. Inf. Syst. Front. **13**, 101–113 (2010)
4. JS Pistole, Fraudulent Identification Documents and the Implications for Homeland Security. *Statement Rec Before House Sel Comm Homel Secur.* 2003. URL: https://www.fbi.gov/news/testimony/fraudulent-identification-documents-and-the-implications-for-homeland-security
5. J Jonas, Identity resolution: 23 years of practical experience and observations at scale, in *Proc 2006 ACM SIGMOD Int Conf Manag data - SIGMOD'06* (ACM Press, New York, NY, USA, 2006), p. 718 [*SIGMOD'06*]
6. GA Wang, H Chen, H Atabakhsh, Automatically detecting deceptive criminal identities. Commun. ACM **47**, 70–76 (2004)
7. GA Wang, HC Chen, JJ Xu, H Atabakhsh, Automatically detecting criminal identity deception: an adaptive detection algorithm. IEEE Trans. Syst. Man. Cybern. Part a-Systems Humans **36**, 988–999 (2006)
8. I Bhattacharya, L Getoor, Collective entity resolution in relational data. ACM Trans. Knowl. Discov. from Data **1**, 5 (2007)
9. H Pasula, B Marthi, B Milch, S Russell, I Shpitser, Identity uncertainty and citation matching. Adv Neural Inf Process Syst 2003:1425–1432.
10. T Donaldson, A position paper on collaborative deceit, in *AAAI-94 Work Plan Interagent Commun*, 1994
11. IP Fellegi, AB Sunter, A theory for record linkage. J. Am. Stat. Assoc. **64**, 1183–1210 (1969)
12. M Bilenko, R Mooney, W Cohen, P Ravikumar, S Fienberg, Adaptive name matching in information integration. IEEE Intell. Syst. **18**, 16–23 (2003)
13. Hernandez MA, Stolfo SJ: The Merge/purge Problem for Large Databases. In Proc 1995 ACM SIGMOD Int Conf Manag data (SIGMOD '95) Edited by Carey MJ, Schneider DA. San Jose, CA; 1995:127–138
14. AE Monge, Matching algorithms within a duplicate detection system. IEEE. Data Eng. Bull **23**, 14–20 (2000)
15. AK Elmagarmid, PG Ipeirotis, VS Verykios, Duplicate record detection: a survey. IEEE Trans. Knowl. Data Eng. **19**, 1–16 (2007)
16. JM Cheek, SR Briggs, Self-consciousness and aspects of identity. J. Res. Pers. **16**, 401–408 (1982)
17. H Tajfel, JC Turner, *The Social Identity Theory of Inter-Group Behavior* (Nelson-Hall, Chicago, 1986)
18. JC Turner, *Some Current Issues in Research on Social Identity and Self-Categorization Theories* (Blackwell, Oxford, 1999)
19. K Deaux, D Martin, Interpersonal networks and social categories: specifying levels of context in identity processes. Soc. Psychol. Q. **66**, 101–117 (2003)
20. S Stryker, RT Serpe, *Commitment, Identity Salience, and Role Behavior: Theory and Research Example* (Springer-Verlap, New York, 1982)
21. TC Redman, The impact of poor data quality on the typical enterprises. Commun. ACM **41**, 79–82 (1998)
22. United Kingdom Home Office: Identity Fraud: A Study. 2002. URL: http://www.homeoffice.gov.uk/cpd/id_fraud-report.pdf
23. R Ananthakrishna, S Chaudhuri, V Ganti, Eliminating Fuzzy Duplicates in Data Warehouses. In *Proc 28th Int Conf Very Large Data Bases*. Hong Kong, China; 2002:586–597.
24. D V Kalashnikov, S Mehrotra, Z Chen, Exploiting relationships for domain-independent data cleaning. In *Proc 2005 SIAM Int Conf Data Min*. Newport Beach, CA; 2005:262-273
25. H Köpcke, E Rahm, Frameworks for entity matching: a comparison. Data Knowl. Eng. **69**, 197–210 (2010)
26. B Marshall, S Kaza, J Xu, H Atabakhsh, T Petersen, C Violette, H Chen, Cross-Jurisdictional Criminal Activity Networks to Support Border and Transportation Security. In *Proceedings 7th Int IEEE Conf Intell Transp Syst*. Washington, D.C.; 2004:100–105.
27. DE Brown, SC Hagen, Data association methods with applications to law enforcement. Decis. Support Syst. **34**, 369–378 (2003)
28. D Dey, S Sarkar, P De, A distance-based approach to entity reconciliation in heterogeneous databases. IEEE Trans. Knowl. Data. Eng. **14**, 567–582 (2002)
29. P Ravikumar, WW Cohen, A Hierarchical Graphical Model for Record Linkage, in *Proceeding 20th Conf Uncertain Artif Intell* (AUAI Press, Arlington, VA, 2004)
30. W. E. Winkler. *Methods for record linkage and Bayesian networks*. Technical Report, Statistical Research Division, U.S. Census Bureau, Washington, DC, 2002.
31. K Nigam, AK McCallum, S Thrun, T Mitchell, Text Classification from Labeled and Unlabeled Documents using EM. Mach. Learn **39**, 103–134 (2000)
32. A Culotta, A McCallum, Joint deduplication of multiple record types in relational data, in *Proc 14th ACM Int Conf Inf Knowl Manag* (ACM, Bremen, Germany, 2005), pp. 257–258
33. I Bhattacharya, L Getoor, Entity resolution in graphs, in *Min graph data* (Wiley-Blackwell, Hoboken, 2006), p. 311
34. S Bartunov, A Korshunov, S Park, W Ryu, H Lee, Joint Link-Attribute User Identity Resolution in Online Social Networks. In *Proc 6th Work Soc Netw Min Anal (SNA-KDD '12)*. Beijing, China; 2012
35. P Hsiung, A Moore, D Neill, J Schneider, Alias detection in link data sets. In *Proc Int Conf Intell Anal*. McLean, VA; 2005.
36. J Xu, GA Wang, J Li, M Chau, Complex problem solving: identity matching based on social contextual information. J. Assoc. Inf. Syst. **8**, 525–545 (2007)
37. MA Jaro, *UNIMATCH: A Record Linkage System: User's Manual*. Washington: The Bureau, 1978
38. HB Newcombe, JM Kennendy, SJ Axford, AP James, Automatic linkage of vital records. Science **130**, 954–959 (1959)
39. J Li, GA Wang, Criminal identity resolution using social behavior and relationship attributes. In *Proc 2011 IEEE Int Conf Intell Secur Informatics, ISI 2011*; 2011:173–175.
40. VI Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady **10**, 707–710 (1966)
41. J Kubica, A Moore, D Cohn, J Schneider, Finding underlying connections: a fast graph-based method for link analysis and collaboration queries. Proc. Int. Conf. Mach. Learn. **20**, 392–399 (2003)
42. J Kubica, AW Moore, D Cohn, J Schneider, cGraph: A fast graph-based method for link analysis and queries. In *Proc 2003 IJCAI Text-Mining Link-Analysis Work;* 2003:22–31