Disclosing false identity through hybrid link analysis

Boongoen, Tossapon; Shen, Qiang; Price, Christopher John

Published in:
Artificial Intelligence and Law

DOI:
10.1007/s10506-010-9085-9

Publication date:
2010

Citation for published version (APA):
Boongoen, T., Shen, Q., & Price, C. J. (2010). Disclosing false identity through hybrid link analysis. *Artificial Intelligence and Law*, *18*(1), 77-102. https://doi.org/10.1007/s10506-010-9085-9

Noname manuscript No.
(will be inserted by the editor)

Disclosing False Identity through Hybrid Link Analysis

Abstract Combating the identity problem is crucial and urgent as false identity has become a common denominator of many serious crimes, including mafia trafficking and terrorism. Without correct identification, it is very difficult for law enforcement authority to intervene, or even trace terrorists' activities. Amongst several identity attributes, personal names are commonly, and effortlessly, falsified or aliased by most criminals. Typical approaches to detecting the use of false identity rely on the similarity measure of textual and other content-based characteristics, which are usually not applicable in the case of highly deceptive, erroneous and unknown descriptions. This barrier can be overcome through analysis of link information displayed by the individual in communication behaviours, financial interactions and social networks. In particular, this paper presents a novel link-based approach that improves existing techniques by integrating multiple link properties in the process of similarity evaluation. It is utilised in a hybrid model that proficiently combines both text-based and link-based measures of examined names to refine the justification of their similarity. This approach is experimentally evaluated against other link-based and text-based techniques, over a terrorist-related dataset, with further generalization to a similar problem occurring in publication databases. The empirical study demonstrates the great potential of this work towards developing an effective identity verification system.

Keywords false identity detection · hybrid algorithm · link analysis · terrorist data

1 Introduction

False identity has become a common denominator of many serious crimes such as mafia trafficking, fraud and money laundering. Particularly in the UK, financial losses due to such cause are reported to be around 1.3 billion pounds each year (Wang et al 2006). Holders of false identity intend to avoid accountability and to leave no traces for law enforcement authority. Identity fraud is intentionally committed with a view to perpetrating another crime from the most trivial to the most dreadful imaginable. Organized criminals make use of counterfeit identity to cover up illicit activities and illicitly gained capital. Especially in the case of terrorism, it is widely utilized to provide

financial and logistical support to terrorist networks that have set up and encourage criminal activities to undermine civil society. Tracking and preventing terrorist activities undoubtedly requires authentic identification of criminals and terrorists who typically possess multiple fraudulent names, dates of birth, addresses, bank accounts, telephone numbers and email accounts.

With present high-quality off-the-shelf equipment, it is easy to generate credible false identity documents. On the other hand, it requires a great deal of time and experience to distinguish between genuine and forged copies. Usually, it is not feasible for a common person to recognize the ten or fifteen security features presented in a document in a short period. However, successful detection can prevent serious consequences such as the September 11 terrorist attacks. In that particular tragic case, US authorities failed to discover the use of false identities by nineteen terrorists, who were all able to enter the United States without any problem. Most of them typically possessed several dates of birth and multiple aliases (Boongoen and Shen 2008b). For instance, 'Mohamed Atta', alleged ringleader of the September 11 attacks, has exploited eight different aliases of 'Mehan Atta', 'Mohammad El Amir', 'Muhammad Atta, Mohamed El Sayed', 'Mohamed Elsayed', 'Muhammad Al Amir Awag Al Sayyid Atta' and 'Muhammad Al Amir Awad Al Sayad'. In such circumstance, identity verification and name variation detection systems (Bilenko and Mooney 2003; Branting 2003; Torvik et al 2004; Wang et al 2006) that rely solely on the inexact search of textual attributes are effective to some extent. Nevertheless, these methods will fail to disclose the truth that highly deceptive identities (e.g 'Usama bin Laden' and 'The prince') refer to the same person (Hsiung et al 2005).

The aforementioned dilemma may be overcome through link analysis, which seeks to discover knowledge based on the relationships in data about people, places, things, and events. Intuitively, despite using distinct false identities, each terrorist normally exhibits unique relations with other entities involved in legitimate activities found in any open or modern society – making use of mobile phones, public transportation and financial systems. Link analysis techniques have proven effective for identity problems (Badia and Kantardzic 2005; Boongoen and Shen 2008b; Hsiung et al 2005; Pantel 2006) by exploiting link information instead of content-based information, which is typically unreliable due to intentional deception, translation and data-entry errors (Wang et al 2005, 2006). Recently, link analysis has also been employed by Argentine intelligence organizations for analyzing Iranian-Embassy telephone records. This specific investigation aims to make a circumstantial case that the Iranian Embassy had been involved in the July 18, 1994 terror bombing of a Jewish community centre (Porter Jan 25, 2008). In addition, this methodology has also been adopted to establish a semantic-based retrieval mechanism on the citation network of legal cases (Zhang and Koppaka 2007) and the spatial criminal network, by which relations amongst co-defendants can be identified (Oatley et al 2005).

To justify the similarity between entities (e.g. names, publications and web pages) in a link network, many well-known algorithms like SimRank (Jeh and Widom 2002), PageSim (Lin et al 2006), Connected-Triple (Reuther and Walter 2006) and Jaccard (Liben-Nowell and Kleinberg 2007) concentrate only on the cardinality of joint neighbors to which they are directly linked. Despite their notable performance, other characteristics of a link pattern have so far been excluded from the underlying analysis. As such, the quality of the similarity evaluation may be enhanced by including the uniqueness measure (Boongoen and Shen 2008b) of an overlapping neighbor context.

Inspired by such insight, this paper presents a novel link-based similarity algorithm, *Connected-Path*, in which multiple link properties are proficiently blended to refine the process of similarity estimation. Also, by following this initial development, a hybrid method for false identity detection is introduced as an intelligent aggregation of the Connected-Path and text-based measures. Unlike the supervised model in (Hsiung et al 2005) that also includes both text-based and link-based metrics, the proposed approach is unsupervised and so avoids the problems of a supervised methodology: unintentional encoding of human bias and noise into training data, scalability to large data collections, and adaptability to new cases. A similar unsupervised method (Angheluta and Moens 2007) resolves cross-document co-references, by incorporating appearances and textual properties (i.e. syntactics and semantics) with link information. That technique relies heavily on a priori linguistic knowledge. As a result, it may be inapplicable to a new problem domain where such information is not available. By contrast, the approach described in this paper is language-independent and knowledge-free, and so can be easily adopted to new problem domains.

The rest of this paper is organized as follows. Section 2 introduces fundamental concepts and practices of false identity detection, upon which the present research is based. Section 3 illustrates the link analysis approach to the identity problem and a number of link-based similarity measures. Following that, the new link-based similarity algorithm, with the underlying path-based intuition, is presented in Section 4. Then, the fifth section includes details of the hybrid method for false identity detection and its core terminology. The experimental evaluation over the terrorist-related and publication data collections is provided in Section 6. The paper is concluded in Section 7, with the perspective of further work.

2 False Identity Detection

Identity is a set of characteristic descriptors unique to a specific person, which can be principally categorized into three types of identity indicators (Clarke 1994; Wang et al 2006):

- attributed identity
- biographical identity
- biometric identity

Attributed identity consists of descriptions of name, details of parents, date and place of birth, and is often used as the primary means of establishing identity. Biographical identity is constituted from personal information over a life span (e.g. criminal, educational and financial history) and can also be exploited for the same purpose. Biometric identity consists of personal measurements such as fingerprints and DNA features.

Attributed and biographical identity indicators are greatly subject to deception as they are much easier to falsify than biometric indicators. The main focus of the current research is to disclose the possibility of attributed identity being falsely or deceptively specified, especially for the case of personal names.

According to Fig 1 which is obtained from the study of identity deception in (Wang et al 2004, 2006), name deception is the most common practice found in the collection of investigated criminal records – 100 percent of occurrence in all 372 cases examined, in

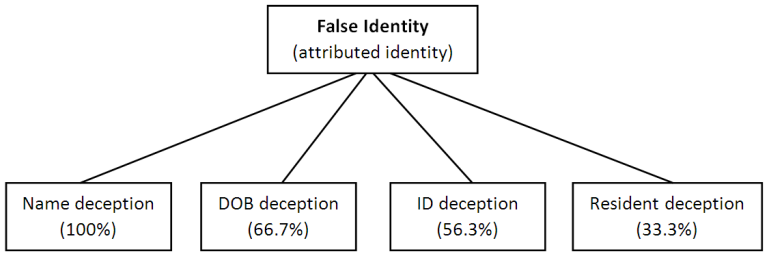

Fig. 1 Taxonomy of attributed identity deception obtained from (Wang et al 2004, 2006). Each percentage number represents the proportion of examined cases that contain a particular deception type.

fact. This set of records involves 24 criminals, each of which possesses one real identity and several deceptive records.

In particular, such illegal acts can be accomplished by employing a combination of the following falsified formats:

- Partly deceptive name, either false first or family name.
- A completely different name.
- Abbreviation or add-on to first or last name.
- Similar pronunciation, but with different spelling.
- Name swap via transposing first and last names.

To battle false identity, an exact-match query to a law enforcement computer system is simply ineffective. A better approach that has been extensively studied in (Bilenko and Mooney 2003; Branting 2003; Torvik et al 2004; Wang et al 2006) is to exploit the similarity measure of names obtained from one or several string-matching techniques. In practice, to measure the similarity $s(p, q)$ of names p and q, the simplest and most common approach is the Levenshtein distance, which is sometimes called simply the Edit distance (Navarro 2001). The distance is defined as a number of edit operations (e.g. character insertion, deletion and substitution) that convert p to q. Note that the greater the Edit distance is, the less similar two names are deemed to be. In addition, an adaptive string matching method has been introduced in (Bilenko et al 2003) as an extension of the conventional metric. To enhance the quality of such distance measure, each of the underlying edit operations is assigned a weight, which is automatically learned from examples.

Following this pioneer method, several string-matching techniques have been devised to handle different forms of name ambiguity: Monge-Elkan, Jaccard, Soundex, Smith-Waterman and q-grams (see more details in (Navarro 2001)). Amongst these, Jaro (Jaro 1995) and q-grams (Kukich 1992), primarily utilised with short strings (e.g. personal names) (Navarro 2001), are widely recognized for their distinguished performance in the area of record-linkage (Fellegi and Sunter 1969). As a result, the Jaro and q-grams similarity measures are employed in the current research to illustrate the performance of the *content-based* approach to identity problems. Despite their reported success in the literature, the aforementioned methods may be ineffective for cases where highly deceptive names are deployed. For instance, they would fail to recognize the association between the following pairs of terrorists' names, whose overlapping textual content is very small, or even nil. Note that these name pairs are obtained from the Terrorist data collection (Hsiung et al 2005), see further details in Section 6.1.

– ‘Ashraf refaat nabith henin’ and ‘Salem ali’
– ‘Fahid mohammed ali msalam’ and ‘Usama al-kini’
– ‘Fadil abdallah muhamad’ and ‘Harun fazul’
– ‘Usama bin laden’ and ‘The prince’
– ‘Usama bin laden’ and ‘The emir’
– ‘Abu mohammed nur al-deen’ and ‘The doctor’

To overcome this limitation, the *link-based* approach has been proposed, taking into account the relations amongst examined names. Many existing link-based methods (Badia and Kantardzic 2005; Boongoen and Shen 2008b; Hsiung et al 2005; Pantel 2006) to false identity detection have employed the intuition that a person (including criminal and terrorist) naturally possesses a unique pattern of relations to other information entities, such as vehicles, bank cards, telephone numbers, email accounts and friends. Thus, false identity may be discovered using a link-based similarity measure which is estimated over the link network of the kind presented in Fig 2. Given such a graphical representation of intelligence data that entails the activities of suspected identities, it is possible to hypothesize that ‘Identity A’ and ‘Identity B’ may actually refer to the same person. This hypothesis emerges since these two identities have frequently used identical email addresses and cash cards.

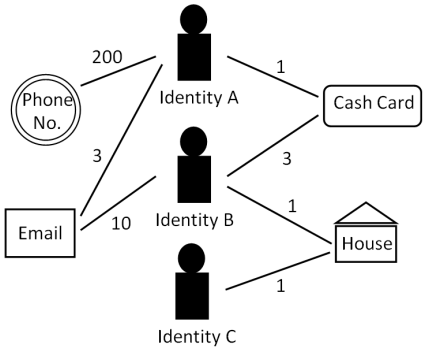

Fig. 2 A link network of intelligence data, in which different identities can be related by identical objects (e.g. email accounts, telephone numbers, accommodation, social groups, cash and credit cards). Note that each number denotes the frequency that a specific pair of entities relate (e.g. co-occur in an observed event).

This idea resembles the methodology of link analysis (Getoor and Diehl 2005; Liben-Nowell and Kleinberg 2007) which has proven effective for a wide range of application domains. For instance, an author-collaboration graph has been employed for personal name resolution in publication databases (Reuther and Walter 2006; Sun et al 2005). Also, given a web graph that represents web pages and their hyper-link relations, the link-based methods of (Hou and Zhang 2003; Lin et al 2006) are used to identify similar web pages. Furthermore, link analysis has also been successful for personal name resolution in an email collection (Minkov et al 2006) and relational entity resolution (Bhattacharya and Getoor 2007). Recently, semantic-association discovery techniques have been introduced to identify conflict-of-interest relationships and integrate social networks (Aleman-Meza et al 2008).

It is noteworthy that, in order to handle all possible formats of name deception, both text-based and link-based similarity measures are included in the supervised model introduced in (Hsiung et al 2005). In spite of high performance illustrated therein, the major drawbacks of this methodology are: (i) inaccuracy of manually marked training data, to which human bias, error or noise may be unintentionally amended, (ii) difficult and laborious scaling up to a large data collection, and (iii) inability to adapt to new cases. By contrast, the hybrid method described herein overcomes the aforementioned barriers by employing an unsupervised approach that intelligently integrates the link-based and text-based metrics to justify the similarity amongst investigated identities.

3 Link Analysis Approach to False Identity Detection

3.1 Problem Formulation

With the link analysis methodology, detecting false identities or aliases is conducted on a link network that represents the relations amongst references (i.e. names) of real-world entities. This graphical scheme has been effectively exploited for the problem of link prediction in (Liben-Nowell and Kleinberg 2007; Murata and Moriyasu 2008) and that of author-name resolution in (Reuther and Walter 2006). Formally, a link network can be specified as an undirected graph $G = (V, W)$. It is represented by two information sets, the set of vertices V and that of weighted edges W. Let X be the set of all references and R be the set of their relations in an examined dataset. Each vertex $v_i \in V$ denotes a specific reference $x_i \in X$. Then, an edge $w_{i,j} \in W$ (linking vertices $v_i, v_j \in V$) corresponds to a relation $r_{i,j} \in R$ between references $x_i, x_j \in X$.

This research concentrates on analyzing a link network whose edges correspond to 'co-occurrence' relations amongst references. In other words, a relation $r_{i,j} \in R$ stands for the fact that references $x_i, x_j \in X$ appear together in a specific observation. It is bi-directional such that $r_{i,j}$ is equivalent to $r_{j,i}, \forall r_{i,j}, r_{j,i} \in R$ and $\forall x_i, x_j \in X$. As a result, edges are undirected and any $w_{i,j}, w_{j,i} \in W$ are inherently equivalent. Thus, the developed paradigm is simple (semantics-free) and efficient regarding its information acquisition and analysis. It can be effectively extended to the highly semantics-embedded case, where both direction and type of examined relations are exhibited within a directed graph (e.g. the semantic network of email communication (Minkov et al 2006) and the citation network of scientific publications (Pasula et al 2003)).

In an undirected graph G, each edge $w_{i,j} \in W$ possesses statistical information $|w_{i,j}| \in \{1, \ldots, \infty\}$, which signifies the frequency of the corresponding relation $r_{i,j} \in R$ (i.e. the frequency of which references x_i and x_j co-occur in the given dataset). By representing the multiplicity of each edge as a frequency count (or weight), the resulting graph terminology becomes simple (i.e. no parallel edges), without losing any potential link information (Wasserman and Faust 1994). Let O be the set of real-world entities each of which is referred to by at least one member of the set X. Any set of two references (x_i, x_j) is an alias pair when both references correspond to the same real-world entity: $(x_i \equiv o_k) \wedge (x_j \equiv o_k), o_k \in O$. In practice, disclosing an alias pair in a graph G involves finding a couple of vertices (v_i, v_j), whose similarity $s(v_i, v_j)$ is significantly high. Intuitively, the higher $s(v_i, v_j)$ is, the more similar vertices v_i and v_j are and hence, the greater the possibility that the corresponding references x_i and x_j, constitute an actual alias pair.

To illustrate this framework, the link network of publication data, similar to that of (Reuther and Walter 2006), is discussed here. A set of author references (i.e. names) and their relations can be presented as a graph in Fig. 3, where $X = \{A, B, C, D, E\}$, $R = \{r_{A,B}, r_{A,C}, r_{A,D}, r_{B,E}, r_{C,D}\}$, and $r_{i,j}$ denotes the fact that references x_i and x_j co-occur as authors of a specific publication. In addition, the edge $w_{A,D}$ is presented with $|w_{A,D}| = 2$ since references A and D are co-authors of two different papers (i.e. $paper_2$ and $paper_3$). Likewise, the frequency statistics $|w_{A,C}|$ of the edge $w_{A,C}$ is 1 as references A and C have only one joint publication, $paper_2$. Given O as the set of real-world author entities, a pair of references, such as (A, E), may be hypothesized, based on their similarity, as the alias pair (i.e. $(A \equiv o_k) \wedge (E \equiv o_k), o_k \in O$).

(a) relations in database table

publications	author names
$paper_1$	A and B
$paper_2$	A, C and D
$paper_3$	A and D
$paper_4$	B and E
$paper_5$	C and D

(b) relations in a graph format

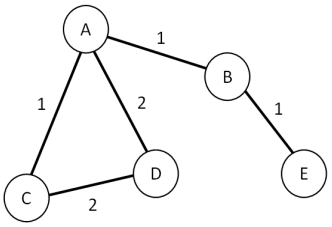

Fig. 3 Relations between author references and publications, presented in: (a) database table format and (b) graph format.

3.2 Link Based Similarity Measures

Unlike the content-based approach in which appearances and other linguistic features of the underlying references are directly compared, link analysis makes use of a link-based similarity that is measured using the link pattern of any pair of references in question. Based on this perspective and the increasing volume of network-like information (such as online resources, publication repository, phone and credit-card usages), several link-based similarity methods have been introduced to evaluate the similarity between various information objects: Co-citation (Small 1973), Jaccard (Liben-Nowell and Kleinberg 2007), SimRank (Jeh and Widom 2002), Connected-Triple (Reuther and Walter 2006), PageSim (Lin et al 2006) and a variety of random walk algorithms (Fouss et al 2007; Minkov et al 2006). See further details in (Getoor and Diehl 2005; Liben-Nowell and Kleinberg 2007).

Many existing link-based similarity measures have concentrated exclusively on the numerical count of shared neighbors. Jeh and Widom (2002) specifically emphasized that 'similar objects are usually linked to similar neighboring objects'. Let e be an entity of interest (e.g. a author reference in a publication network) and N_e be a set of entities that are directly linked to e, called neighbors of e. The similarity between entities e_1 and e_2 is then determined by the cardinality of their common neighbors $|N_{e_1} \cap N_{e_2}|$, where N_{e_1} and N_{e_2} are the set of neighbors of entity e_1 and that of e_2, respectively. In essence, the higher the cardinality is, the greater the similarity of these entities becomes (Liben-Nowell and Kleinberg 2007).

This basic concept has been adopted by the co-citation (Small 1973) and several other well-known methods that are used to reveal the interesting relationships amongst scientific publications. For instance, the Connected-Triple technique (Reuther and Walter 2006) evaluates the similarity of objects given their overlapping social context. It is exploited to disclose duplicated author references in a publication database, which is represented as a social network $G = (V, W)$. In particular, author references and their co-author relations are represented by the set of vertices V and a set of edges W, respectively. The similarity of any two vertices $v_i, v_j \in V$ can be estimated by counting the number of Connected-Triples they are part of. Formally, a Connected Triple, $Triple = \{V_{Triple}, W_{Triple}\}$, is a subgraph of G containing three vertices $V_{Triple} = \{v_i, v_j, v_k\} \subset V$ and two edges $W_{Triple} = \{w_{i,k}, w_{j,k}\} \subset W$, with $w_{i,j} \notin W$. Fig 4 presents an example of a social network in which object a and object b are considered similar due to the fact that there exist two Connected-Triples connecting them together, $V_{Triple1} = \{a, b, c\}, W_{Triple1} = \{w_{a,c}, w_{b,c}\}$ and $V_{Triple2} = \{a, b, d\}, W_{Triple2} = \{w_{a,d}, w_{b,d}\}$, with $w_{a,b} \notin W$.

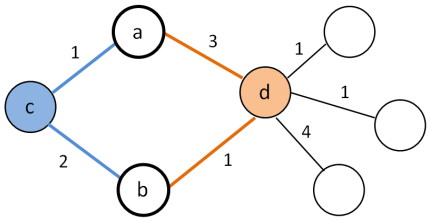

Fig. 4 Example of a social network with Connected-Triples.

Despite their simplicity, the aforementioned cardinality based approaches are greatly sensitive to noise and often generate a large proportion of false positives (Klink et al 2006). This shortcoming emerges because these methods exclusively concern the cardinality aspect of link patterns without taking into account another link property. As the first attempt to extend this approach, Boongoen and Shen (2008b) suggested the *uniqueness measure* as the additional criterion to the cardinality in order to crystalize the estimation of similarity values. The resulting mechanism proved effective over the terrorism-related data collection (Hsiung et al 2005).

Given a graph $G = (V, W)$ in which objects and their relations are members of the set of vertices V and those of the set of edges W, respectively, a uniqueness measure $UQ^k_{i,j}$ of any two vertices $v_i, v_j \in V$ can be approximated from each joint neighbor $v_k \in V, w_{i,k}, w_{j,k} \in W$ such that:

$$UQ^k_{i,j} = \frac{|w_{i,k}| + |w_{j,k}|}{\sum\limits_{\forall v_m \in V} |w_{m,k}|} \tag{1}$$

Here, $|w_{i,k}|$ is the weight (i.e. frequency) of the edge between vertices $v_i, v_k \in V$, $|w_{j,k}|$ is the weight of the edge between vertices v_j and v_k, and $|w_{m,k}|$ is the weight of the edge between v_k and any other vertex $v_m \in V$.

To summarize the uniqueness measures of joint link patterns $UQ_{i,j}$ between vertices v_i and v_j, the ratios estimated for each shared neighbor are aggregated as

$$UQ_{i,j} = \frac{1}{\beta} \sum_{k=1}^{\beta} UQ_{i,j}^{k} \tag{2}$$

where β is the number of those overlapping neighbors that vertices v_i and v_j are commonly linked to, i.e. $\beta = |N_{v_i} \cap N_{v_j}|$. With the example given earlier in Fig. 4, the uniqueness measure $UQ_{a,b}$ between vertices a and b can be estimated as

$$UQ_{a,b} = \frac{1}{2}(UQ_{a,b}^{c} + UQ_{a,b}^{d}) = \frac{1}{2}(\frac{3}{3} + \frac{4}{10})$$

4 A New Link-Based Similarity Algorithm: Connected-Path

This section presents a novel link-based similarity method, Connected-Path, that employs multiple properties of a link pattern (i.e. cardinality and uniqueness) for estimating a degree of similarity. In particular, this new path-based algorithm is established upon the simple practice of counting a number of shared neighbors that are employed by many existing techniques (Liben-Nowell and Kleinberg 2007; Reuther and Walter 2006). However, it takes into account the neighboring context more widely than the adjacent span originally studied.

Following the terminology of a link network $G = (V, W)$ given in Section 3.1, a path between two vertices $v_i, v_j \in V$, $path(v_i, v_j)$, is a sequence of unique vertices $\{v_i, v_1, \ldots, v_n, v_j\}$ such that edges $w_{i,1}, w_{1,2}, \ldots, w_{n-1,n}, w_{n,j} \in W$. The length of path p is $length(p) = |p| - 1$, where $|p|$ is the number of vertices in path p. In addition, $PATH(v_i, v_j, r)$ denotes the set of all possible paths between vertices $v_i, v_j \in V$, whose length satisfies the condition $2 \leq length(p) \leq r$. Analogous to the Connected-Triple algorithm, a direct path $p^*, length(p^*) = 1$, between any two vertices v_i and v_j is not considered. This is due to the fact that such a path does not represent the environment in which two vertices co-occur and its inclusion may lead to an incorrect conclusion.

The similarity between vertices $v_i, v_j \in V$ is determined by the accumulated uniqueness measure that is obtained from all paths $p \in PATH(v_i, v_j, r)$. This measure can be formally defined as follows:

$$Connected - Path(v_i, v_j) = \sum_{p \in PATH(v_i, v_j, r)} \frac{U(p)}{length(p)} \tag{3}$$

where $U(p)$ is the uniqueness of path p, which is calculated using Equation 4. Note that the path uniqueness is divided by its length, as longer paths are intuitively considered to be less informative than shorter ones.

$$U(path(v_i, v_j)) = \prod_{v_x \in path(v_i, v_j), v_x \notin \{v_i, v_j\}} UQ(v_x) \tag{4}$$

Here, $UQ(v_x)$ is the uniqueness score measured at the vertex $v_x \in path(v_i, v_j)$, which can be estimated using Equation 5. The multiplication is specifically used in order to summarize the compositional quality of paths and magnify their differentiation (i.e. contribution towards a similarity estimate).

$$UQ(v_x) = \frac{|w_{x,x-1}| + |w_{x,x+1}|}{\sum\limits_{\forall v_g \in V} |w_{x,g}|} \tag{5}$$

Here, an edge between the vertex $v_x \in path(v_i, v_j)$ and any other vertex $v_g \in V$ is denoted as $w_{x,g}$, and $w_{x,x-1}$ and $w_{x,x+1}$ represent edges from v_x to its adjacent vertices in the path, $v_{x-1}, v_{x+1} \in path(v_i, v_j)$.

Effectively, the similarity measure $S_{Connected-Path}(v_i, v_j) \in [0, 1]$ between vertices $v_i, v_j \in V$ is obtained by the following normalization, where $Connected - Path_{max}$ is the maximum estimate between any two vertices in a link network G. Note that this process is necessary for the hybrid model, in which both link-based and text-based measures are represented on a unified scale.

$$S_{Connected-Path}(v_i, v_j) = \frac{Connected - Path(v_i, v_j)}{Connected - Path_{max}} \tag{6}$$

By applying this methodology to the task of false identity detection, the higher the similarity measure, the greater the possibility that names represented by vertices v_i and v_j constitute the use of false identity. It is noteworthy that longer paths (i.e. higher value of r) make the overall estimation more refined and robust, but at the cost of greater computational requirements. Fig. 5 depicts an example of different neighborhood scopes that may be included in the similarity estimation. Note that the network is extracted from the Terrorist dataset (Hsiung et al 2005), where each vertex corresponds to a specific name of terrorist or event (given in the bracket). The simplest variation of the Connected-Path method explores only the short paths, whose $length = 2$. Each of these paths includes a unique adjacent common-neighbor of the two vertices in question. Particularly to the path $p_x = \{v_i, v_a, v_j\}$, v_i and v_j (i.e. 'The teacher' and 'Usama bin laden') can be considered similar as they are linked to the joint neighbor v_a (i.e. 'Abu abdallah'). In other words, 'The teacher' and 'Usama bin laden' may be aliases of the same person, provided that these two co-occur with another name 'Abu abdallah' in the reported news. This conforms the intuition taken by a number of link-based similarity techniques, Connected-Triple (Reuther and Walter 2006) and SimRank (Jeh and Widom 2002), for instance.

Despite its efficiency, the aforementioned model evaluates the similarity based on information at a rather coarse level. The underlying measure can be refined by expanding the scope of shared neighbors beyond those adjacent ones. This means taking into account longer paths (i.e. $length > 2$), e.g. the path $p_y = \{v_i, v_b, v_c, v_j\}$ shown in Fig 5. Intuitively, v_b and v_c (i.e. 'Abu fatima' and 'Abu muhammad') that co-occur in the collected data, provide a common neighboring context with which v_i and v_j (i.e. 'The prince' and 'Usama bin laden') associate. Hence, the similarity between v_i and v_j can be better estimated through more remote neighbors along such paths, in addition to exploiting associations with their immediate neighbors.

The same concept is applicable to the path $p_z = \{v_i, v_d, v_e, v_f, v_j\}$. Despite the fact that they do not co-occur, v_d and v_f (i.e. 'Ayman al-zawahiri' and 'Muhammad atef') are similar as they relate to the same event of '8/7/98 US embassy bombings', i.e. v_e. This similarity can be propagated through the link pattern such that v_i and v_j can be justified alike. However, as compared to a short path p_x in this example, longer paths p_y and p_z provides less informative and implicit evidence for justifying the similarity between v_i and v_j. Unlike p_x that accounts for a primary, identical common-neighbour, p_y and p_z include only so-called 'secondary' common-neighbours, which are

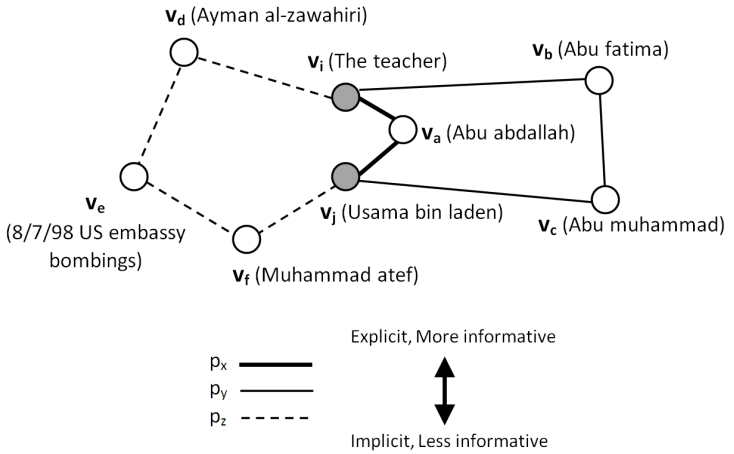

Fig. 5 Example of link paths between vertices v_i, v_j, each with a different path length: $length(p_x) = 2, lenght(p_y) = 3$ and $length(p_z) = 4$.

not identical but considered to be similar given their shared link pattern. The strength of each common neighbor, primary or secondary, decays by the path length, which is reflected in Equation 3. It is noteworthy that a number of existing link-based similarity algorithms, e.g. SimRank (Jeh and Widom 2002) and PageSim (Lin et al 2006), adopt an analogous practice of investigating joint neighbors that may be several edges away from a given pair of vertices under examination.

5 Intelligent Hybrid Method to False Identity Detection

In order to achieve an identity resolution system that is capable of disclosing various deception types (including the use of totally different names), this section presents an unsupervised hybrid approach that combines both text-based and link-based similarity measures through a re-ranking mechanism. The underlying intuition is illustrated in Fig. 6. For a set of names X where each name $x_i \in X$ is represented as a vertex $v_i \in V$ in a link network $G = (V, W)$, a set of highly possible alias pairs, (x_i, x_j) (equivalently represented as (v_i, v_j)), are identified as follows:

- *Step1*: Generate a collection of φ pair-wise link-based similarity degrees $s(v_i, v_j)$ for all pairs of $v_i, v_j \in V$, using a link-based method. That is, $\varphi = \frac{N(N-1)}{2}$, where N denotes the number of vertices in the link network G. In particular, for the Connected-Path algorithm that has been proposed in the preceding section, $s(v_i, v_j) = S_{Connected-Path}(v_i, v_j)$.
- *Step2*: Arrange these pair-wise measures in descending order of their magnitude, and select the first K pairs that are of the top values. Note that $K << \varphi$ is pre-defined, allowing subjective input of intelligence analysts to be incorporated.
- *Step3*: Estimate the final similarity $s^*(x_i, x_j)$ of each selected name pair (x_i, x_j) (represented by (v_i, v_j) in the link network) by aggregating its link-based and text-based similarity measures:

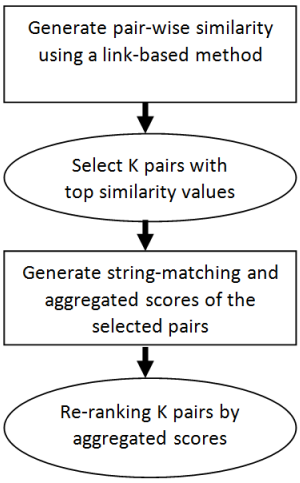

Fig. 6 Descriptive model of the hybrid approach.

$$s^*(x_i, x_j) = AGG\left(s(v_i, v_j), str(x_i, x_j)\right) \tag{7}$$

where $str(x_i, x_j)$ is the text-based similarity score that can be obtained using a string-matching technique like that of (Jaro 1995). Here, AGG denotes an aggregation operator that is employed to combine the link-based and text-based measures. For the present research, the following four aggregation models are investigated:

(i) *AGG-Text*, where s^* is defined by

$$s^*(x_i, x_j) = str(x_i, x_j) \tag{8}$$

(ii) *AGG-Average*, where s^* is defined by

$$s^*(x_i, x_j) = \frac{s(v_i, v_j) + str(x_i, x_j)}{2} \tag{9}$$

(iii) *AGG-Max*, where s^* is defined by

$$s^*(x_i, x_j) = \max\left(s(v_i, v_j), str(x_i, x_j)\right) \tag{10}$$

(iv) *AGG-Min*, where s^* is defined by

$$s^*(x_i, x_j) = \min\left(s(v_i, v_j), str(x_i, x_j)\right) \tag{11}$$

– *Step4*: Identify the likely alias pairs in accordance with the assumption that the higher the similarity $s^*(x_i, x_j)$ is, the greater such possibility becomes.

Effectively, the proposed hybrid model does not require the manual preparation of training samples (i.e. supervised instances), in which human bias and error are usually encoded. Such a process can be laborious or even infeasible with a large data collection. Another distinct advantage of this unsupervised method is its ability to adapt to new

cases, where the supervised counterpart normally fails. This is particularly the case as a supervised model usually relies on a limited number of historical scenarios or rules that encompass some but not all possible different combinations of analytic variables.

6 Empirical Evaluation

In order to evaluate the performance of the proposed approach, similarity estimates acquired from Connected-Path and the hybrid method are compared with those derived by other link-based and text-based techniques, for the task of discovering aliases in the terrorism-related dataset (Hsiung et al 2005). Further assessment is also carried out with a publication data collection (Reuther and Walter 2006).

6.1 Investigated Datasets

Terrorist. To reflect the difficulty of detecting aliases in intelligence data, this dataset has been constructed by extracting 919 real alias pairs from terrorism-related web pages and news stories (Hsiung et al 2005). Each of the 4,088 nodes in this link network corresponds to a name of person (criminal/terrorist), place, organization or event, while each of the 5,581 links denotes the co-occurrence of a specific pair of names with its weight representing the frequency of such occurrence. Fig. 7 shows an example of this link network in which the names 'Bin laden' and 'Abu abdallah' truly refer to the same real-world person. Note that the model originally developed for this dataset is not the same as that used here, due to their fundamental differences regarding the adopted learning schemes: supervised and unsupervised, respectively.

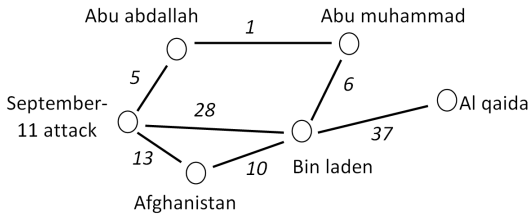

Fig. 7 Example of the Terrorist dataset.

DBLP1-2. In order to evaluate the generality of the Connected-Path algorithm and the hybrid model, additional experiments are carried out to identify duplicated author names in the DBLP (Digital Bibliography and Library Project) publication data collection (Reuther and Walter 2006). With the same terminology and format, there are originally three versions of such dataset: DBLP-SUB01, DBLP-SUB02 and DBLP-SUB03. For the present work, applications to DBLP-SUB01 and DBLP-SUB02 are examined, with the corresponding datasets abbreviated to DBLP1 and DBLP2 hereafter. Statistically, DBLP1 consists of 2,796 author names, 8,157 co-authoring links and 23 duplicated name pairs. DBLP2 is larger than the first version with 6,351 names, 18,543 links and 73 duplicated name pairs. Fig. 8 depicts an example of this dataset where 'Hepu Deng' and 'Hepu Dong' are references to the same author.

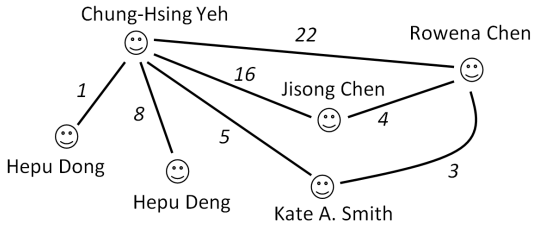

Fig. 8 An example of the DBLP dataset.

Note that aliases in the Terrorist dataset are caused mostly by deception and translation errors. As shown in Fig. 9, about 70 percent of true alias pairs are difficult to recognize – with the Jaro string-matching scores (Jaro 1995) being less than 0.6. Approximately 20 percent of them are highly deceptive, with a nil matching degree. On the other hand, duplicates in the DBLP data collection are brought about mainly by data entry errors. In particular, around 65-70 percent of duplicated name pairs possess very high Jaro scores (more than 0.8). Although this problem is not subject to deception, the DBLP1-2 datasets are included to support performance evaluation between the proposed model and other comparable methods.

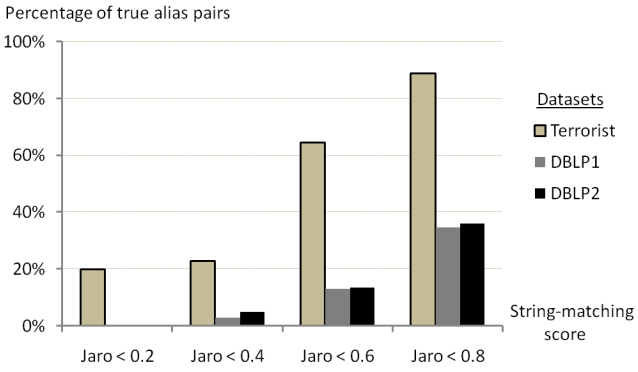

Fig. 9 Percentage of true alias/duplicated pairs in the Terrorist and DBLP1-2 datasets, categorized in accordance with their string-matching scores, i.e. Jaro measures (Jaro 1995).

6.2 Compared Methods

The performance of the Connected-Path method and the hybrid model are assessed against many state-of-the-art link-based similarity algorithms and string-matching measures. These compared methods include

– *Connected Triple (CT)*: This link-based technique was originally used to reveal possible duplicated names in the DBLP data collection (Reuther and Walter 2006). Its core concept is provided earlier in Section 3.2.

– *Jaccard (JC)*: Given a graph $G = (V, W)$, this metric is commonly used in information retrieval (Liben-Nowell and Kleinberg 2007) and measures the similarity $s(v_i, v_j)$ between any two vertices $v_i, v_j \in V$ by

$$s(v_i, v_j) = \frac{|N_{v_i} \cap N_{v_j}|}{|N_{v_i} \cup N_{v_j}|} \tag{12}$$

where $N_{v_x} \subset V$ denotes the set of neighbors of $v_x \in V$, i.e. $\forall v_y \in N_{v_x}, w_{x,y} \in W$. A similar metric, called 'interest distance', has been developed to deduce shared-interest relations between people based on the history of email communication (Schwartz and Wood 1993). This distance measure $d(v_i, v_j)$ between vertices $v_i, v_j \in V$ is defined by

$$d(v_i, v_j) = \frac{|N_{v_i} \cup N_{v_j}| - |N_{v_i} \cap N_{v_j}|}{|N_{v_i} \cup N_{v_j}|} \tag{13}$$

Note that this is equivalent to the Jaccard coefficient, where $s(v_i, v_j) = 1 - d(v_i, v_j)$.

– *Pointwise Mutual Information (PMI)*: This metric has been extended and applied to estimating the similarity amongst vertices in a link network (Pantel 2006). Given a graph $G = (V, W)$, a frequency vector $F_{v_i} = (f_{i,1}, f_{i,2}, \ldots, f_{i,|V|})$ is constructed for each vertex $v_i \in V$, where $|V|$ is the total number of vertices and $f_{i,j}, j = 1 \ldots |V|$ is defined as follows:

$$f_{i,j} = \begin{cases} w_{i,j} & if\ w_{i,j} \in W \\ 0 & otherwise \end{cases} \tag{14}$$

The mutual information vector $M_{v_i} = (m_{i,1}, m_{i,2}, \ldots, m_{i,|V|})$ is created for each vertex $v_i \in V$, where $m_{i,j}$ is the pointwise mutual information between vertices v_i, v_j that are estimated by

$$m_{i,j} = \log \frac{\frac{f_{i,j}}{\alpha}}{\frac{\sum_{x=1}^{|V|} f_{i,x}}{\alpha} \times \frac{\sum_{y=1}^{|V|} f_{y,j}}{\alpha}} \tag{15}$$

where α is the total frequency count in the given graph G, i.e. $\alpha = \sum_{x=1}^{|V|-1} \sum_{y=x+1}^{|V|} f_{x,y}$. Following that, the similarity $s(v_i, v_j)$ between any two vertices $v_i, v_j \in V$ can be found using the cosine coefficient of their mutual information vectors (i.e. M_{v_i} and M_{v_j}), which is given by

$$s(v_i, v_j) = \frac{\sum_{\forall x} m_{i,x} \times m_{j,x}}{\sqrt{\sum_{\forall y} m_{i,y}^2 \times \sum_{\forall y} m_{y,j}^2}} \tag{16}$$

– *SimRank (SR)*: With the principal objective of finding similar publications given their citation relations, the SimRank algorithm (Jeh and Widom 2002) relies on the cardinality of shared neighbors that are iteratively refined to a fixed point. In each iteration, the similarity $s(v_i, v_j)$ of any pair of vertices $v_i, v_j \in V$ is approximated as follows:

$$s(v_i, v_j) = \frac{C \sum_{p=1}^{|N_{v_i}|} \sum_{q=1}^{|N_{v_j}|} s(N_{v_i}^p, N_{v_j}^q)}{|N_{v_i}||N_{v_j}|} \tag{17}$$

where $N_{v_i}, N_{v_j} \subset V$ are sets of neighboring vertices to which vertices v_i and v_j are linked, respectively. Individual neighbors of both vertices are denoted as $N_{v_i}^p$ and $N_{v_j}^q$, for $1 \leq p \leq |N_{v_i}|$ and $1 \leq q \leq |N_{v_j}|$. The constant $C \in [0, 1]$ is a decay factor that represents the confidence level of accepting two non-identical entities as similar. Note that $s(v_i, v_j) = 0$ when $N_{v_i} = \emptyset$ or $N_{v_j} = \emptyset$.

– *PageSim (PS)*: Within a different domain, PageSim (Lin et al 2006) was developed to capture similar web pages based on associations implied by their hyperlinks. In essence, a similarity $s(v_i, v_j)$ of vertices v_i and v_j is dictated by the coherence of ranking scores $R(v_g, v_i)$ and $R(v_g, v_j)$ propagated to them from any other vertex $v_g \in V$. It is noteworthy that ranking scores are explicitly generated using the page ranking scheme, PageRank (Brin and Page 1998), of the well-known Google search engine.

Given a link network $G = (V, W)$, let $P(v_i)$ denote the PageRank score of a vertex $v_i \in V$. Its value can be estimated from the following iterative refinement (i.e. $P(v_i) = \lim_{k \to \infty} P_k(v_i)$):

$$P_k(v_i) = (1 - \beta) + \beta \sum_{v_j \in V, w_{i,j} \in W} P_{k-1}(v_j) \times Dist(v_j, v_i) \tag{18}$$

where β is a dampening factor that is usually set to 0.85 (Brin and Page 1998), and $P_0(v_i), \forall v_i \in V$ is initially set to 1. In addition, $Dist(v_j, v_i)$ can be found by

$$Dist(v_j, v_i) = \frac{|w_{j,i}|}{\sum_{v_x \in V, v_x \neq v_j} |w_{j,x}|} \tag{19}$$

Having achieved this, the score $R(v_i, v_j)$ propagated from $v_i \in V$ to $v_j \in V$ can be calculated as follows:

$$R(v_i, v_j) = \sum_{p \in PATH(v_i, v_j, r)} d \times P(v_i) \times PDist(p, v_i, v_j) \tag{20}$$

where $d \in (0, 1]$ is a decay factor, r is the maximum path length, and $PDist(p, v_i, v_j)$ is defined by the following equation with v_{x+1} denoting the vertex adjacent to v_x in path p, along the direction from v_i to v_j:

$$PDist(p, v_i, v_j) = \prod_{v_x \in p, v_x \neq v_j} \frac{|w_{x,x+1}|}{\sum_{v_y \in V, v_y \neq v_x} |w_{x,y}|} \tag{21}$$

Effectively, the similarity measure can be defined as

$$s(v_i, v_j) = \sum_{\forall v_g \in V, v_g \notin \{v_i, v_j\}} \frac{\min(R(v_g, v_i), R(v_g, v_j))^2}{\max(R(v_g, v_i), R(v_g, v_j))} \tag{22}$$

– *Jaro (JR)*: In addition to the aforementioned link-based methods, the Jaro string-matching measure (Jaro 1995) is employed to illustrate the effectiveness of the text-based approach and also to set the base-line performance for the underlying tasks. This distance-based metric relies on the number and order of the common characters between strings $p = a_1 \dots a_K$ and $q = b_1 \dots b_L$. Particularly, a character $a_i \in p$ is *common with* q when there is $b_j \in q, b_j = a_i$ such that $i - H \leq j \leq i + H$ and

$$H = \frac{\min(|p|, |q|)}{2} \tag{23}$$

where $|p|$ denotes the length of string p.

Let $p' = a'_1 \dots a'_{K^*}$, $(K^* \leq K)$ be the sequence of characters in p that are common with q (in the same order that they appear in p) and $q' = b'_1 \dots b'_{L^*}$, $(L^* \leq L)$ be the similar sequence of characters in q that are common with p, *transposition for* p' and q' is specified as the cardinality of positions i where $a'_i \neq b'_i$. Effectively, the Jaro similarity metric for strings p and q is defined as follows.

$$Jaro(p, q) = \frac{1}{3}\left(\frac{|p'|}{|p|} + \frac{|q'|}{|q|} + \frac{|p'| - T_{p',q'}}{|p'|}\right) \tag{24}$$

where $T_{p',q'}$ is a half of transposition for p' and q'.

– *q-grams (QG)*: To consolidate the underlying evaluation, the performance of the q-grams (also called n-grams) string matching method (Kukich 1992) is also assessed. q-grams are substrings of length q in longer strings. Note that bigrams ($q = 2$) are exploited in the current evaluation, while other commonly used q-grams are unigrams ($q = 1$) and trigrams ($q = 3$). For instance, 'peter' contains the bigrams 'pe', 'et', 'te' and 'er'. A bigram similarity measure between two strings is estimated by counting the number of bigrams in common (i.e. bigrams contained in both strings) and dividing the count by the number of bigrams in the shorter string (called 'Overlap coefficient'). Other alternatives are the number in the longer string (called 'Jaccard similarity') and the average number of q-grams in both strings (called the 'Dice coefficient').

6.3 Experimental Results

6.3.1 Performance of the Connected-Path Method

Table 1 shows the number of alias pairs that are disclosed by each examined method, with respect to the number of retrieved entity pairs ($K \in \{200, 400, 600, 800, 1000\}$) of the highest similarity values in each dataset (Terrorist, DBLP1, DBLP2). Note that 'CP' denotes the Connected-Path technique and the corresponding 'Precision' and 'Recall' measures are estimated by

$$Precision = \frac{Number\ of\ disclosed\ alias\ pairs}{Number\ of\ retrieved\ entity\ pairs} \tag{25}$$

$$Recall = \frac{Number\ of\ disclosed\ alias\ pairs}{Number\ of\ all\ alias\ pairs} \tag{26}$$

Table 1 Number of alias pairs that are discovered in the Terrorist and DBLP1-2 datasets by each method, where K is the number of retrieved pairs of the highest similarity values. The corresponding (Precision/Recall) measures are given in brackets.

Dataset	Method	Number of K				
		200	400	600	800	1,000
Terrorist	CP (r = 4)	52 (0.26/0.06)	81 (0.20/0.09)	136 (0.23/0.15)	170 (0.21/0.18)	193 (0.19/0.21)
	CP (r = 2)	18 (0.09/0.02)	76 (0.19/0.08)	98 (0.16/0.11)	140 (0.18/0.15)	165 (0.17/0.18)
	CT	1 (0.01/0.00)	5 (0.01/0.01)	5 (0.01/0.01)	77 (0.10/0.08)	77 (0.08/0.08)
	JC	7 (0.04/0.01)	15 (0.04/0.02)	23 (0.04/0.03)	29 (0.04/0.03)	37 (0.04/0.04)
	PMI	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)
	SR	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)
	PS	7 (0.04/0.01)	36 (0.09/0.04)	63 (0.11/0.07)	79 (0.10/0.09)	92 (0.09/0.10)
	JR	22 (0.11/0.02)	33 (0.08/0.03)	40 (0.06/0.04)	43 (0.05/0.05)	47 (0.05/0.06)
	QG	21 (0.11/0.02)	31 (0.07/0.03)	37 (0.06/0.04)	46 (0.06/0.05)	53 (0.05/0.06)
DBLP1	CP (r=4)	5 (0.03/0.22)	6 (0.02/0.26)	10 (0.02/0.43)	11 (0.01/0.48)	11 (0.01/0.48)
	CP (r=2)	3 (0.01/0.13)	4 (0.01/0.17)	9 (0.01/0.39)	10 (0.01/0.43)	11 (0.01/0.48)
	CT	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)
	JC	2 (0.01/0.09)	3 (0.01/0.13)	7 (0.01/0.30)	7 (0.01/0.30)	8 (0.01/0.35)
	PMI	0 (0.0/0.0)	3 (0.01/0.13)	5 (0.01/0.22)	7 (0.01/0.30)	9 (0.01/0.39)
	SR	1 (0.01/0.04)	2 (0.01/0.09)	3 (0.01/0.13)	3 (0.00/0.13)	3 (0.00/0.13)
	PS	1 (0.01/0.04)	2 (0.01/0.09)	4 (0.01/0.17)	4 (0.01/0.17)	4 (0.00/0.17)
	JR	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)
	QG	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)
DBLP2	CP (r=4)	4 (0.02/0.05)	9 (0.02/0.12)	14 (0.02/0.19)	16 (0.02/0.22)	20 (0.02/0.27)
	CP (r=2)	3 (0.02/0.04)	8 (0.02/0.11)	12 (0.02/0.16)	15 (0.02/0.21)	18 (0.02/0.25)
	CT	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)
	JC	3 (0.02/0.04)	8 (0.02/0.11)	8 (0.01/0.11)	8 (0.01/0.11)	8 (0.01/0.11)
	PMI	2 (0.01/0.03)	4 (0.01/0.05)	4 (0.01/0.05)	5 (0.01/0.07)	6 (0.01/0.08)
	SR	1 (0.01/0.01)	1 (0.00/0.01)	2 (0.00/0.03)	3 (0.00/0.04)	3 (0.00/0.04)
	PS	1 (0.01/0.01)	1 (0.00/0.01)	1 (0.00/0.01)	4 (0.01/0.05)	5 (0.01/0.07)
	JR	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)
	QG	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)

These results show that the new Connected-Path algorithm (CP) consistently outperforms string-matching techniques (JR and QG) and link-based similarity algorithms (CT, JC, PMI, SR and PS), over the identified sets of top-K similar entity pairs. Also, the more-refined CP model (i.e. $r = 4$) achieves a better performance than its basic counterpart (i.e. $r = 2$). This finding indicates that long paths can help to improve the underlying similarity measure, thereby reinforcing the quality of the resulting CP method. Amongst the link-based methods investigated, PS possesses the best precision/recall statistics. Analogous to the CP approach, PS has such a superior outcome because of the inclusion of edge weights (i.e. co-occurrence frequencies) in its similarity estimation process. On the other hand, CT is ineffective as it does not take account of this link property.

Unlike the CP and CT techniques that concentrate on joint neighbors, JC, PMI and SR also consider unshared neighbors that may reduce the strength of the similarity measure. For the Terrorist network which contains a large number of uninformative edges (i.e. arbitrary and semantic-free co-occurrences), such methods become inaccurate. Their performance degrades even more when the similarity evaluation is carried out beyond the set of adjacent neighbors. However, they perform better (as compared to their application to Terrorist) on the DBLP1-2 datasets in which links (i.e. co-authoring relations) are more reliable. In spite of its low performance, the SR measure, which has been recognized as a benchmark for link-based analysis technique for the publication (Getoor and Diehl 2005) and Internet (Calado et al 2006) domains, is included in the present evaluation to reflect the difficulty of deceptive alias detection. Of course, as false identity detection is an extremely difficult task, it is generally the case that precision/recall statistics are much lower than what might be expected in usual classification problems. Despite this, the proposed method leads to the best overall performance, for a variety of data collections.

Note that the string-matching algorithms are effective in discovering a minority of alias pairs in the Terrorist dataset: those with a very high appearance-based similarity. However, these methods become ineffective with 'highly deceptive' cases where overlapping textual content is extremely small, or even nil. Based on the collection of 183 name pairs in the Terrorist data that are highly deceptive (i.e. whose JR measures are 0), Fig. 10 presents the number of such pairs that can be revealed per link-based method. These results demonstrate that the CP approach is effective for tackling the deception problem, with its performance being generally robust to different parameter (r) settings. The results of PMI and SR are omitted since they are totally ineffective at detecting any of the deceptive pairs.

6.3.2 Effectiveness of the Hybrid Model

For comparison purposes, the following steps are employed to evaluate the performance of the proposed hybrid model for false identity detection:

- *Step1*: The Connected-Path method is used to derive pair-wise link-based similarity $S_{Connected-Path}(v_i, v_j)$ for any $v_i, v_j \in V$, where each $v_i \in V$ corresponds to a particular name $x_i \in X$.
- *Step2*: $K = 1,000$ name pairs (e.g. equivalently represented as (v_i, v_j)) with top Connected-Path measures are selected.
- *Step3*: These selected pairs are re-ranked in accordance with their ultimate similarity score $s^*(x_i, x_j)$, using each of the aggregation models introduced in Section 5:

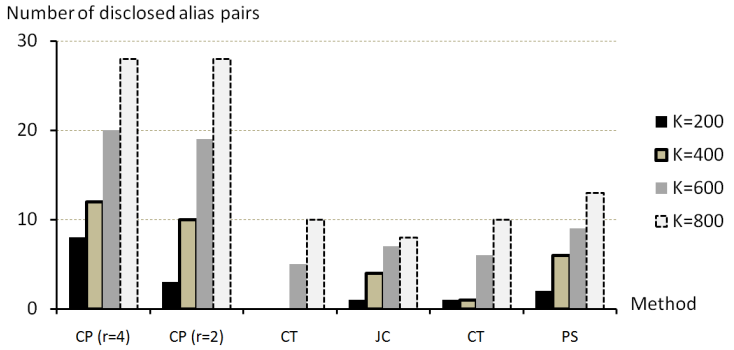

Fig. 10 Number of 'highly deceptive' alias pairs in the Terrorist dataset that can be discovered from the K name pairs of the highest similarity measures.

AGG-Text, AGG-Average, AGG-Max and AGG-Min. In particular, the Jaro technique (JR) is utilized to generate the string-matching scores of those selected name pairs.

The results of Table 2 show that the hybrid models can further improve the performance achievable by the Connected-Path ($r = 4$) method alone. The number of false positives is substantially reduced, especially regarding the collections of top-200 and top-400 name pairs. The most and the least effective hybrid methods are AGG-Text and AGG-Min, respectively. These findings which can be analogously observed on all investigated datasets, strongly indicate that the proposed hybrid approach is robust to the choice of aggregation mechanism that is used to combine the text-based and link-based similarity measures. Note that this positive performance is also observed when the less accurate, but more efficient, Connected-Path estimation (i.e. $r = 2$) is used in the hybrid models. The corresponding results on the Terrorist and DBLP1-2 datasets are given in Figs 11-13.

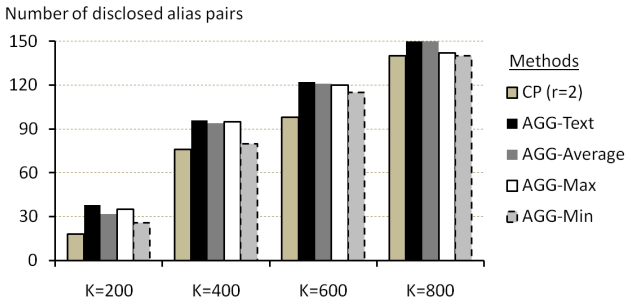

Fig. 11 Number of alias pairs discovered in the Terrorist dataset by different hybrid models, where K is the number of retrieved pairs which are of the highest similarity values. Note that CP($r = 2$) is employed to generate the link-based similarity measures in these hybrid methods.

Table 2 Number of alias pairs discovered in the Terrorist and DBLP1-2 datasets by different hybrid models, where K is the number of retrieved pairs which are of the highest similarity values. Note that CP($r = 4$) is employed to generate the link-based similarity measures in these hybrid methods.

Dataset	Method	Number of K				
		200	400	600	800	1,000
Terrorist	CP(r = 4)	52	81	136	170	193
	AGG-Text	81	123	149	175	193
	AGG-Average	75	122	148	171	193
	AGG-Max	79	123	149	170	193
	AGG-Min	57	106	144	170	193
DBLP1	CP(r = 4)	5	6	10	11	11
	AGG-Text	10	11	11	11	11
	AGG-Average	9	10	11	11	11
	AGG-Max	7	8	11	11	11
	AGG-Min	6	7	10	11	11
DBLP2	CP(r = 4)	4	9	14	16	20
	AGG-Text	18	20	20	20	20
	AGG-Average	17	19	20	20	20
	AGG-Max	14	16	19	20	20
	AGG-Min	10	13	15	18	20

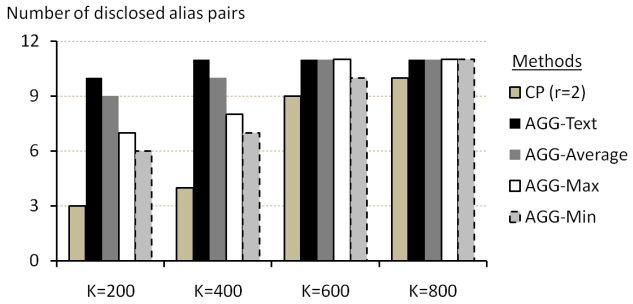

Fig. 12 Number of alias pairs discovered in the DBLP1 dataset by different hybrid models, where K is the number of retrieved pairs which are of the highest similarity values. Note that CP($r = 2$) is employed to generate the link-based similarity measures in these hybrid methods.

6.4 Complexity Analysis

6.4.1 Complexity of Connected-Path and Compared Link-Based Algorithms

In addition to evaluating these methods in terms of discovered alias pairs, it is important to investigate the computational complexity that would determine their actual efficiency for real-world applications. Suppose that a link network consists of n distinct entities (i.e. vertices), each averagely linked to m other entities. The time complexity for the Connected-Path method to generate all pair-wise similarity values is $O(n^2m^r)$, where r is the maximum length of paths that are included in the similarity estimation and $r \in \{2, \ldots, \infty\}$.

CT is the most efficient amongst the compared link-based methods, with its time complexity being $O(n^2m)$. Both JC and PMI are slightly more expensive than the

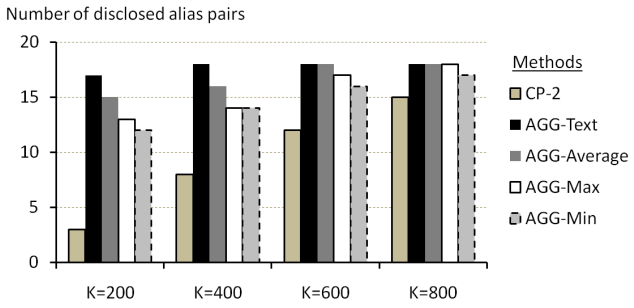

Fig. 13 Number of alias pairs discovered in the DBLP2 dataset by different hybrid models, where K is the number of retrieved pairs which are of the highest similarity values. Note that CP($r = 2$) is employed to generate the link-based similarity measures in these hybrid methods.

simple CT model, where their complexity is generally $O(n^2m^2)$. With the f iterations of refinement, the time complexity of SimRank algorithm is $O(n^2m^2f)$. Note that the results shown in Table 1 are obtained using $f = 3$ (with its usual range being around 3-5). In contrast, PageSim is rather more complex compared to the others as it begins with ranking all entities using the PageRank technique, whose time complexity is $O(nmk)$ where k is the number of iterations for refining the ranking values ($k = 3$ in this experiment). Having accomplished the ranking process, the similarity of two entities is estimated on the ranking values propagated from their shared neighbors, with the maximum connecting-path length of r (r and d set to 3 and 0.8 for the results given in Table 1). Hence, the overall time complexity of PageSim method is $O(n^2m^{2r} + nmt)$.

It is noteworthy that the outcomes of the Connected-Path measure given in Table 1 are achieved by setting the maximum path length as $r = 4$, which consequently results in the time complexity of $O(n^2m^4)$. Essentially, this requirement can be substantially reduced to $O(n^2m^2)$ by including only short paths (i.e. $r = 2$), in which case the number of disclosed aliases drops slightly, but it is still considerably larger than those of its counterparts. These results imply the efficient exploitation and flexibility of the Connected-Path algorithm in real-time applications.

6.4.2 Efficiency of the Hybrid Approach

The time complexity of the proposed hybrid model is the combination of those required by the underlying string-matching (i.e. Jaro) and Connected-Path methods. In general, the time complexity of Jaro is $O(l^2)$, where l denotes the average length of studied names. Thus, the hybrid method with Connected-Path possesses the overall time complexity of $O(n^2m^r + Kl^2)$. This is reduced to $O(n^2m^2 + Kl^2)$ when the shortest-path variation of Connected-Path ($r = 2$) is employed. It is $O(n^2m^4 + Kl^2)$ if $r = 4$.

The above analysis indicates that the hybrid method offers great flexibility in its potential applications. The algorithm specification can vary with respect to different time requirements. In particular, the Connected-Path measure that includes only short paths is efficient for a quick or immediate response, whilst the measure with longer paths can be carried out in the background, as the supporting module for the former.

7 Conclusion

This paper has presented a new link-based similarity method, Connected-Path, that exploits multiple link properties to estimate the degree of similarity. Unlike many existing link-based algorithms that concentrate exclusively on the cardinality of joint neighbors, this measure further includes the uniqueness property of the link patterns in order to refine the similarity estimate. Connected-Path outperforms both well-known link-based and text-based methods, on the terrorist-related and publication data collections. Furthermore, the paper has offered a hybrid model for name disambiguation, which efficiently aggregates both link-based and text-based similarity metrics. It has been empirically demonstrated that this hybrid approach can enhance the performance obtained by using the link-based measure alone.

Despite such achievements, their efficacy may be further demonstrated with more terrorist-related or similar intelligence datasets. In addition, the proposed novel link-based similarity measure can be exploited for resolving identities and aggregating relevant scenarios in the environment of intelligence data analysis (Shen et al 2006). Specifically, as part of on-going work, the concepts of qualitative reasoning have been adopted to enhance a conventional numerical link analysis (that usually fails to achieve accurate and coherent interpretation of similarity measures). Based on the order-of-magnitude model (Raiman 1991), an initial qualitative method of (Boongoen and Shen 2009a) is introduced such that the similarity and associated link properties can be expressed by linguistic descriptors. Effectively, it allows the detection results to be naturally explained and validated. This is similar to the computational model of (Ashley and Bruninghaus 2009), which aims to classify case texts and predict case outcomes in a manner that can be explained in terms that law practitioners can understand.

In spite of its simplicity, a significant limitation exists with the aforementioned order-of-magnitude based model. This is because of its ineffective interpretation of the underlying real-valued variables and ambiguous products of interval-based qualitative values. In addition, this model does not address the gradual nature of qualitative labels, i.e. the extent to which a suspect's height of 170 cm. is 'moderate' or 'tall' is often a matter of degree and differently perceived by one analyst or another (Ali et al 2003; Shen and Leitch 1992). To improve this initial method, the theory of fuzzy sets (Zadeh 1965) may be employed to represent the qualitative model such that the vagueness and uncertainty inherent in human knowledge and judgement can be better captured and rationalized. For AI and Law research, the potential of such practice has long been recognized. For instance, it has provided a mathematical means to link the determinacy of decisions to fact patterns described in indeterminate terms (Philipps and Sartor 1999).

Another on-going research project concentrates on combining multiple link properties using the methodology of OWA (Ordered Weighted Averaging) (Boongoen and Shen 2008a). The preliminary method of (Boongoen and Shen 2009b) makes use of stress functions (Yager 2007), by which users can determine the actual behavior of an aggregation process. Although the current work uses only a limited number of predefined stress functions, it is in line with the attempt to bridge law practitioners and AI researchers (Oskamp and Lauritsen 2002). Indeed, this aggregation model allows practicing lawyers to guide the formulation of computational methods at high levels without necessarily making choices about the underlying mathematical details.

Acknowledgements This work is sponsored by UK EPSRC grant EP/D057086. The authors are grateful to the members of the project team for their contribution, whilst taking full responsibility for the views expressed in this paper. The authors would also like to thank the anonymous referees for their constructive comments which have helped considerably in revising this work.

References

Aleman-Meza B, Nagarajan M, Ding L, Sheth AP, Arpinar IB, Joshi A, Finin T (2008) Scalable semantic analytics on social networks for addressing the problem of conflict of interest detection. ACM Transactions on the Web 2(1):1–29

Ali AH, Dubois D, Prade H (2003) Qualitative reasoning based on fuzzy relative orders of magnitude. IEEE Transactions on Fuzzy Systems 11(1):9–23

Angheluta R, Moens MF (2007) Cross-document entity tracking. In: Proceedings of European Conference on IR Research, pp 670–673

Ashley KD, Bruninghaus S (2009) Automatically classifying case texts and predicting outcomes. Artificial Intelligence and Law 17:125–165

Badia A, Kantardzic MM (2005) Link analysis tools for intelligence and counterterrorism. In: Proceedings of IEEE International Conference on Intelligence and Security Informatics, Atlanta, pp 49–59

Bhattacharya I, Getoor L (2007) Collective entity resolution in relational data. ACM Transactions on KDD 1(1)

Bilenko M, Mooney RJ (2003) Adaptive duplicate detection using learnable string similarity measures. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 39–48

Bilenko M, Mooney R, Cohen W, Ravikumar P, Fienberg S (2003) Adaptive name matching in information integration. IEEE Intelligent Systems 18(5):16–23

Boongoen T, Shen Q (2008a) Clus-DOWA: A New Dependent OWA Operator. In: Proceedings of IEEE International Conference on Fuzzy Sets and Systems, pp 1057–1063

Boongoen T, Shen Q (2008b) Detecting false identity through behavioural patterns. In: Proceedings of International Crime Science Conference, London

Boongoen T, Shen Q (2009a) Order-of-magnitude based link analysis for false identity detection. In: Proceedings of the 23rd International Workshop on Qualitative Reasoning, pp 7–15

Boongoen T, Shen Q (2009b) Semi-Supervised OWA Aggregation for Link-Based Similarity Evaluation and Alias Detection. In: Proceedings of IEEE International Conference on Fuzzy Sets and Systems, pp 288–293

Branting K (2003) A comparative evaluation of name matching algorithms. In: Proceedings of International Conference on AI and Law, pp 224–232

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems 30(1-7):107–117

Calado P, Cristo M, Gonçalves MA, de Moura ES, Ribeiro-Neto BA, Ziviani N (2006) Link based similarity measures for the classification of web documents. Journal of American Society for Information Science and Technology 57(2):208–221

Clarke R (1994) Human identification in information systems: Management challenges and public policy issues. IT and People 7(4):6–37

Fellegi I, Sunter A (1969) Theory of record linkage. Journal of the American Statistical Association 64:1183–1210

Fouss F, Pirotte A, Renders JM, Saerens M (2007) Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Transactions on Knowledge and Data Engineering 19(3):355–369

Getoor L, Diehl CP (2005) Link mining: a survey. ACM SIGKDD Explorations Newsletter 7(2):3–12

Hou J, Zhang Y (2003) Effectively finding relevant web pages from linkage information. IEEE Transactions on Knowledge and Data Engineering 15(4):940–951

Hsiung P, Moore A, Neill D, Schneider J (2005) Alias detection in link data sets. In: Proceedings of International Conference on Intelligence Analysis

Jaro MA (1995) Probabilistic linkage of large public health data files. Statistics in Medicine 14(5-7):491–498

Jeh G, Widom J (2002) Simrank: A measure of structural-context similarity. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 538–543

Klink S, Reuther P, Weber A, Walter B, Ley M (2006) Analysing social networks within bibliographical data. In: Proceedings of International Conference on Database and Expert Systems Applications, Poland, pp 234–243

Kukich K (1992) Techniques for automatically correcting words in text. ACM Computing Surveys 24(4):377–439

Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology 58(7):1019–1031

Lin Z, King I, Lyu MR (2006) Pagesim: A novel link-based similarity measure for the world wide web. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, pp 687–693

Minkov E, Cohen WW, Ng AY (2006) Contextual search and name disambiguation in email using graphs. In: Proceedings of Int. Conference on Research and Development in IR, pp 27–34

Murata T, Moriyasu S (2008) Link prediction based on structural properties of online social networks. New Generation Computing 26:245–257

Navarro G (2001) A guided tour to approximate string matching. ACM Computing Surveys 33(1):31–88

Oatley GC, Zeleznikow J, Ewart BW (2005) Criminal networks and spatial density. In: Proceedings of International Conference on Artificial Intelligence and Law, pp 246–247

Oskamp A, Lauritsen M (2002) AI in law practice? So far, not much. Artificial Intelligence and Law 10:227–236

Pantel P (2006) Alias detection in malicious environments. In: Proceedings of AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection, Washington, D.C., pp 14–20

Pasula H, Marthi B, Milch B, Russell S, Shpitser I (2003) Identity uncertainty and citation matching. Advances in Neural Information Processing Systems 15:1425–1432

Philipps L, Sartor G (1999) Introduction: From legal theories to neural networks and fuzzy reasoning. Artificial Intelligence and Law 7:115–128

Porter G (Jan 25, 2008) Crying (iranian) wolf in argentina. Asia Times Online

Raiman O (1991) Order of magnitude reasoning. Artificial Intelligence 51(1-3):11–38

Reuther P, Walter B (2006) Survey on test collections and techniques for personal name matching. International Journal on Metadata, Semantics and Ontologies 1(2):89–99

Schwartz ME, Wood DCM (1993) Discovering shared interests using graph analysis. Communications of ACM 36(8):78–89

Shen Q, Leitch R (1992) On extending the quantity space in qualitative reasoning. Artificial Intelligence in Engineering 7:167–173

Shen Q, Keppens J, Aitken C, Schafer B, Lee M (2006) A scenario-driven decision support system for serious crime investigation. Law, Probability and Risk 5(2):87–117

Small H (1973) Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science 24:265–269

Sun J, Qu H, Chakrabarti D, Faloutsos C (2005) Relevance search and anomaly detection in bipartite graphs. ACM SIGKDD Explorations Newsletter 7(2):48–55

Torvik V, Weeber M, Swanson DW, Smalheiser NR (2004) A probabilistic similarity metric for medline records: a model of author name disambiguation. Journal of the American Society for Information Science and Technology 56(2):140–158

Wang GA, Chen H, Atabakhsh H (2004) Automatically detecting deceptive criminal identities. Communications of the ACM 47(3):71–76

Wang GA, Atabakhsh H, Petersen T, Chen H (2005) Discovering identity problems: A case study. In: Proceedings of IEEE International Conference on Intelligence and Security Informatics, Atlanta, pp 368–373

Wang GA, Chen H, Xu JJ, Atabakhsh H (2006) Automatically detecting criminal identity deception: an adaptive detection algorithm. IEEE Transactions on Systems, Man and Cybernetics, Part A 36(5):988–999

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press

Yager RR (2007) Using stress functions to obtain OWA operators. IEEE Transactions on Fuzzy Systems 15(6):1122–1129

Zadeh LA (1965) Fuzzy sets. Information and Control 8:338–353

Zhang P, Koppaka L (2007) Semantics-based legal citation network. In: Proceedings of International Conference on Artificial Intelligence and Law, pp 123–130