Fuzzy Order-of-Magnitude Based Link Analysis for Qualitative Alias Detection*

Qiang Shen, Tossapon Boongoen and Chris Price
Department of Computer Science
Aberystwyth University, UK
{qqs,tsb,cjp}@aber.ac.uk

Abstract

Numerical link-based similarity techniques have proven effective for identifying similar objects in the Internet and publication domains. However, for cases involving unduly high similarity measures, these methods usually generate inaccurate results. Also, they are often restricted to measuring over single properties only. This paper presents an order-of-magnitude based similarity mechanism that integrates multiple link properties to derive semantic-rich similarity descriptions. The approach extends conventional order-of-magnitude reasoning with the theory of fuzzy sets. The inherent ability of this work in computing-with-words also allows coherent interpretation and communication within a decision-making group. The proposed approach is applied to supporting the analysis of intelligence data. When evaluated over a difficult terrorism-related dataset, experimental results show that the approach helps to partly resolve the problem of false positives.

1 Introduction

Disclosing aliases or ambiguous references plays an important role in addressing many challenging real-world applications, including: intelligence analysis [Boongoen *et al.*, 2010], [Branting, 2002], [Wang *et al.*, 2006], information retrieval [Baeza and Ribeiro, 1999], record linkage [Fellegi and Sunter, 1969], entity resolution [Bhattacharya and Getoor, 2007], [Kalashnikov and Mehrotra, 2006] and information integration [Bilenko *et al.*, 2003]. Earlier attempts made use of approximate string matching techniques, such as Levenshtein [Navarro, 2001] and its learnable variations [Bilenko and Mooney, 2003], to compare references' textual content and grade their similarity. These are effective for a name variation caused by typographical and translation errors, for instance, 'John Doe' and 'Jon Doe', or 'Mohamed Atta' and 'Muhammad Atta'. However, for intelligence data analysis, text-based methods are inefficient, and may even be misleading, due to falsified descriptions of terrorists' name and contact details. They would fail drastically to reveal the unconventional truth of highly deceptive identity like that between 'Osama bin Laden' and 'The Prince' [Hsiung *et al.*, 2005].

This issue may be resolved through link analysis, which seeks to discover knowledge based on the relationships in data about people, places, things, and events [Getoor and Diehl, 2005]. For example, in counter-terrorism domain, despite employing distinct false names, each terrorist normally exhibits unique relations with other entities involving in legitimate activities found in any open or modern society (e.g. making use of mobile phones, public transportation and financial systems). Given information on entity names and their associations, link analysis can be performed on an undirected graph underlying such relations. Within such a network, a vertex represents a particular name and an edge corresponds to the relationship between two names. Effectively, plausible aliases can be specified as those pairs of vertices of a high similarity degree. This relation-oriented methodology has been successfully adopted for modelling and analysis of various problems: terrorist and criminal organizations [Popp and Yen, 2006], co-author collaboration [Bhattacharya and Getoor, 2007], [Reuther and Walter, 2006], actor-actor relation [Malin *et al.*, 2005], citation network [Pasula *et al.*, 2003], email communication [Hölzer *et al.*, 2005], [Minkov *et al.*, 2006] and online social ties [Adamic and Adar, 2003] (see [Newman, 2003] for survey and examples).

Whilst precise, reasoning with numerical variables which underpins the existing techniques often suffers from the inaccuracy of quantitative modelling. Such inaccurate descriptions may be caused by a few observations with unduly high values. As a result, the measures of other instances are very small and their interpretations may become rather misleading. Qualitative reasoning in general, and the order-of-magnitude models [Raiman, 1991] in particular have drawn significant interest in tackling these problems. In essence, observed measures and their subsequent aggregations are each represented by a qualitative label which denotes the order-of-magnitude of its information content [Travé-Massuyès and Dague, 2003]. Besides, conventional approaches are typically based on measuring over single properties and hence, are computationally inefficient. Inspired by these observations, this paper presents a novel qualitative model of link analysis that is able to handle multiple link properties.

However, as pointed out in [Ali *et al.*, 2003], [Shen and

[*] An earlier version of this paper was published in *IEEE Transactions on Knowledge and Data Engineering.*

Leitch, 1992], [Sugeno and Yasukawa, 1993], pure symbolic calculi and crisp distinctions amongst qualitative descriptors often encounter important limitations in data and knowledge modelling, such as ad-hoc definition of operators and unnatural semantic interpretation. In this work fuzzy sets are incorporated in the qualitative mechanism to provide a formal theoretical framework. The resulting mechanism offers flexible and interpretable semantics of order-of-magnitude labels, while being supported with mathematically-sound operations. Note that recently, the natural connection between link analysis and fuzzy set theory has been similarly recognised in [Yager, 2008]. This granular technology has been exploited for effective information retrieval [Bordogna *et al.*, 2003], [Kraft *et al.*, 1998] amongst many other applications.

Unlike a number of intelligence analysis methods [Hsiung *et al.*, 2005], [Pantel, 2006], [Wang *et al.*, 2006] which follow a supervised approach, *unsupervised link analysis* is implemented here. This is in order to avoid typical difficulties with the supervised paradigm: required construction of a good training set that includes a representative collection of positive and negative examples (which unintentionally encodes human bias and noise into training data), and limited scalability to large data collections and adaptability to new cases [Bhattacharya and Getoor, 2007]. Interestingly, the unsupervised methodology has also been implemented in [Branting, 2002], where a collection of string comparison algorithms are exploited. Despite the notable accomplishment of that work, it is simply ineffective for highly deceptive cases. In addition, typical flexible information retrieval techniques (e.g. [Bordogna *et al.*, 2003], [Kraft *et al.*, 1998]) will not work well through aggregation of approximate text-based comparisons between a user's query and documents, if applied to cases where deliberately misleading aliases are used. Likewise, many linguistic models for identifying morphologically similar words [Baroni *et al.*, 2002], [Schone and Jurafsky, 2000] that combine string-matching and semantic measures, are also ineffective for intelligence data analysis.

The rest of this paper is organised as follows. Section 2 reviews the basic concepts of absolute order-of-magnitude reasoning, upon which the present research is developed. Section 3 presents the proposed fuzzy-set based approach to interval-valued order-of-magnitude reasoning. This includes the specification of order-of-magnitude descriptors and their fuzzy semantics entailment, the introduction of relevance degrees, and the mechanism for aggregating qualitative information. Section 4 describes the motivation of qualitative link analysis and its implementation using the fuzzy order-of-magnitude model. The assessment of its application to detecting aliases in intelligence data is detailed in Section 5. The paper is concluded in Section 6, with the perspective of further work suggested.

2 Absolute Order-of-Magnitude (AOM) Reasoning

2.1 The AOM Model

The absolute order-of-magnitude model of a qualitative variable V_x operates on a finite set of ordered labels L^x (or qualitative descriptors), which is achieved via a partition of the underlying universe of discourse $U^x \subset \mathcal{R}$ into a set of interval P^x. That is, $P^x = \{p^x_1, \ldots, p^x_{n^x}\}$ and $L^x = \{l^x_1 \ldots l^x_{n^x}\}$, where n^x is the number of intervals/labels and $l^x_1 < \ldots < l^x_{n^x}$ denote the qualitative orders of magnitude. The crisp boundaries of an interval are determined by one or two members of the landmark set $M^x = \{m^x_1, \ldots, m^x_{n^x-1}\}$. Each interval p^x_j is qualitatively expressed by the label $l^x_j, \forall j = 1 \ldots n^x$, and its value range is defined by the lower bound α^x_j and the upper bound β^x_j such that $\alpha^x_j, \beta^x_j \in M^x$ and $\alpha^x_j \leq \beta^x_j$. In essence, landmarks are domain dependent and determined by either justification of human experts or learning from data.

An order-of-magnitude (OM) space S^x defined for a qualitative variable V_x is the combination of the ordered label set L^x and the interval-like treatment of such labels. For example, the value of V_x may be expressed by the set of basic labels $L^x = \{B_1, \ldots, B_{n^x}\}$, with $B_1 < \ldots < B_{n^x}$ denoting the qualitative order amongst the basic labels, meaning that $\alpha < \beta, \forall \alpha \in B_i, \beta \in B_j, i < j$. The corresponding OM space S^x is formally described as $S^x = L^x \cup \{[B_i, B_j] | B_i, B_j \in L^x, i \leq j\}$. That is, the label $[B_i, B_j]$ with $i < j$ is defined as the union of all the elements within the set $\{B_i, B_{i+1}, \ldots, B_j\}$. This representation allows reasoning with single or combined labels to be carried out in the same form.

There is a partial order relation $\leq_p$ in S^x, which can be interpreted as *being more precise than* or *being less general than*. For any labels $Q_p, Q_q \in S^x$, $Q_p \leq_p Q_q$ holds only if $Q_p \subset Q_q$. For easy referencing, $\forall O \in S^x$, the sets $B_O = \{B \in L^x - \{0\}, B \leq_p O\}$ and $B^*_O = \{B \in L^x, B \leq_p O\}$ are termed the *base of O* and the *enlarged base of O*, respectively.

2.2 Qualitative Algebra of AOM

At the outset, the mathematical structure of the AOM model, called 'Qualitative Algebra' or 'Q-algebra', was initially defined as the unification of sign and interval algebra over a continuum of qualitative partitions of the real line [Travé-Massuyès and Piera, 1989]. Subsequently, the notion of 'qualitative expression of a real operator' has been introduced [Agell *et al.*, 2000]. In particular, qualitative operators are considered as multidimensional functions defined in an AOM space. The Cartesian product of $S^1, S^2, \ldots, S^k$ (where value set $S^a, a \in \{1, \ldots, k\}$, corresponds to an examined variable V_a) is adopted to express the outcome of a real operator in $\mathcal{R}^k$ qualitatively, which is later reflected onto the resulting qualitative space $S^{'}$.

Given a real operator ω defined on $\mathcal{R}^k$ involving k real variables $(V_1, \ldots, V_k)$ with each taking values in $\mathcal{R}$, the corresponding qualitative abstraction of ω, denoted as $[\omega]$, is specified as follows:

$$[\omega](X_1, X_2, \ldots, X_k) = [\omega(X_1, X_2, \ldots, X_k)]_{S'} \quad (1)$$

where $X_i \in S^i, i = 1 \ldots k$ is the qualitative label that corresponds to the value of variable V_i and $\omega(X_1, X_2, \ldots, X_k) = \{\omega(x_1, x_2, \ldots, x_k), \forall x_i \in X_i\}$. Inherently, $[\omega]$ assigns to each k-tuple element of $(X_1, X_2, \ldots, X_k)$ a qualitative description of the subset enclosing all underlying numerical results of applying ω over all real values in $X_1, X_2, \ldots, X_k$.

It is feasible to generate the qualitative operator, $[\omega]$, from the uniform ordered labels of an OM space, S, where $S^i = S, i \in \{1, \ldots, k\}$. For any $[\omega]$ and $X_1, X_2, \ldots, X_k \in S$:

$$[\omega](X_1, X_2, \ldots, X_k) = \bigcup_{D_i \in B^*_{X_i}} [\omega](D_1, D_2, \ldots, D_k) \quad (2)$$

According to Equation 1, the qualitative operator $[\omega]$ can be generalised as follows:

$$[\omega](X_1, X_2, \ldots, X_k) = \bigcup_{D_i \in B^*_{X_i}} [\omega(D_1, D_2 \ldots, D_k)]_{S'} \quad (3)$$

To illustrate this algebra, a qualitative extension of the real *sum* operator is used to combine two qualitative labels $Q_1, Q_2 \in S$, whose result is an element belonging to another OM space denoted by $S^{'}$. Note that S and $S^{'}$ are order-of-magnitude spaces defined on $\mathcal{R}$ by the landmark sets $\{-2, -1, 0, 1, 2\}$ and $\{-2, 0, 2\}$, respectively (see Fig 1).

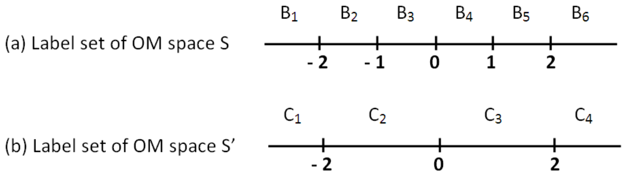

Figure 1: Description of example OM spaces S and $S^{'}$.

Given the labels of Q_1 and Q_2 being $[B_4, B_4]$ and $[B_4, B_5]$, their summation is:

$$\begin{aligned}[sum](Q_1, Q_2) &= \bigcup_{D_i \in B^*_{Q_i}} [sum(D_1, D_2)]_{S'} \\ &= [sum(B_4, B_4)]_{S'} \cup [sum(B_4, B_5)]_{S'}\end{aligned} \quad (4)$$

where $B^*_{Q_1} = \{B_4\}$ and $B^*_{Q_2} = \{B_4, B_5\}$. Based on Fig 1, real values denoted by labels B_4 and B_5 are the intervals $[0, 1)$ and $[1, 2)$, respectively. Hence, the above equation is simplified as

$$\begin{aligned}[sum](Q_1, Q_2) &= [sum([0, 1), [0, 1))]_{S'} \cup [sum([0, 1), [1, 2))]_{S'} \\ &= [[0, 2)]_{S'} \cup [[1, 3)]_{S'}\end{aligned} \quad (5)$$

With the aforementioned label set of OM space $S^{'}$, values within the intervals $[0, 2)$ and $[1, 3)$ are enclosed by the minimal labels of $[C_3, C_3]$ and $[C_3, C_4]$. Then, the result can be achieved as follows:

$$[sum](Q_1, Q_2) = [[C_3, C_3] \cup [C_3, C_4]]_{S'} \ = [C_3, C_4]_{S'} \quad (6)$$

2.3 Homogenisation of References in Multi-Granularity AOM

The $[\omega]$ operator is compatible only to variables specified in the same order-of-magnitude space. To enhance its applicability, this qualitative operator has been extended to multi-granularity domains via a process of homogenisation of references [Agell *et al.*, 2000].

To illustrate this concept, the aforementioned qualitative operator $[sum]$ is employed to aggregate the values of two qualitative variables V_{CT} and V_{UQ}, whose OM spaces (S^{CT} and S^{UQ}) are defined by the landmark sets $M^{CT} = \{2, 6\}$ and $M^{UQ} = \{0.1, 0.3, 0.6, 0.8\}$, respectively. The corresponding label sets are $L^{CT} = \{Small, Medium, Large\}$ and $L^{UQ} = \{Very\ Low, Low, Moderate, High, Very\ High\}$. Suppose that these two variables express the magnitude of two different link properties (see Section 4): CT (Cardinality) and UQ (Uniqueness), where $U^{CT} = [0, \infty)$ and $U^{UQ} = [0, 1]$. Since L^{CT} and L^{UQ} are of unequal granularity, they have to be 'homogenised' onto a common scale, by which references of distinct label sets can be uniformly manipulated and integrated. The homogenisation process is summarised below:

- *Step1*: Sort each landmark set $M^i = \{m^i_1, \ldots, m^i_{n^i-1}\}$ (n^i denotes the number of intervals/labels specified for the variable V_i) into an ascending arrangement, where $m^i_p \leq m^i_q, \forall p < q$. A central landmark $m^i_c \in M^i$ can be specified either as $m^i_{\frac{n^i-1}{2}}$ or $m^i_{\frac{n^i-1}{2}+1}$ when $n^i - 1$ is even, and $m^i_{\frac{n^i}{2}}$ when $n^i - 1$ is odd. Then, each landmark $m^i_t \in M^i$ is translated to the new landmark sm^i_t using $sm^i_t = m^i_t - m^i_c$. Note that the central landmark is now 0 in the new scale. For the example shown in Table 1, the selected central landmarks are $m^{CT}_c = 2$ and $m^{UQ}_c = 0.3$, respectively.
- *Step2*: Landmarks appearing on both sides of 0 (the central landmark) may be dissimilar. A symmetric pattern can be achieved by adding missing landmarks, so that landmarks of the same order-of-magnitude can be found on both sides of 0. These newly added elements are for balancing purpose only; they will not be used to represent values and are deliberately marked as irrelevant. Following the example of Table 1, given the current $M^{CT} = \{0, 4\}$, the additional -4 is appended to M^{CT} such that $M^{CT} = \{-4, 0, 4\}$ and 4 appears on both sides of 0. Similarly, the landmark set M^{UQ} is transformed to $\{-0.5, -0.3, -0.2, 0, 0.2, 0.3, 0.5\}$, where $-0.5, -0.3$ and 0.2 are marked irrelevant.
- *Step3*: Each landmark set is further modified by adding new landmarks on both sides of 0, in such a way that all landmark sets have the same cardinality. Similar to Step 2, new elements are irrelevant with respect to each particular property and are artificially created to support the unification. In the example of Table 1, with the maximum cardinality of M^{CT} and M^{UQ} being 7, additional four landmarks are to be added to the landmark set of M^{CT}, whilst M^{UQ} remains unchanged. Particularly, M^{CT} becomes $\{-4, -2, -1, 0, 1, 2, 4\}$, where $-2, -1, 1$ and 2 are irrelevant landmarks. Using the common OM space S^H with M^H being $\{-3, -2, -1, 0, 1, 2, 3\}$, M^{CT} and M^{UQ} can finally be homogenised by mapping landmarks $\{2 \rightarrow 0, 6 \rightarrow 3\}$ and $\{0.1 \rightarrow -1, 0.3 \rightarrow 0, 0.6 \rightarrow 2, 0.8 \rightarrow 3\}$, respectively.

Table 1: Homogenised landmarks (M^H).

Landmarks	M^{CT}	M^{UQ}
Original	2, 6	0.1, 0.3, 0.6, 0.8
Step1	0, 4	-0.2, 0, 0.3, 0.5
Step2	-4, 0, 4	-0.5, -0.3, -0.2, 0, 0.2, 0.3, 0.5
Step3	-4, -2, -1, 0, 1, 2, 4	-0.5, -0.3, -0.2, 0, 0.2, 0.3, 0.5
Homogenized (M^H)	-3, -2, -1, 0, 1, 2, 3	-3, -2, -1, 0, 1, 2, 3
Irrelevant	-3, -2, -1, 1, 2	-3, -2, 1

Given the homogenised OM space, the qualitative operator $[sum]$ can be used to aggregate the values of qualitative variables V_{CT} and V_{UQ} (see Equations 4-6). Significant limitations exist with this AOM model, however. Firstly, the homogenisation process is difficult to scale up to a large number of input OM spaces. The other problem is due to ineffective interpretation of the underlying real-valued variables and ambiguous products of interval-based qualitative values. This model does not address the gradual nature of qualitative labels, i.e. the extent to which the magnitude of $V_{UQ} = 0.6$ being 'Moderate' or 'High' is often a matter of degree and differently perceived from one analyst to another [Ali *et al.*, 2003], [Shen and Leitch, 1992]. To resolve this problem, the theory of fuzzy sets is applied to represent the qualitative model such that vagueness and imprecision inherent in the underlying human knowledge and judgement can be better captured and rationalised.

3 Fuzzy Set Based Approach to AOM

3.1 Fuzzy AOM Model

In describing a variable V_Z, its value is regarded as a fuzzy set defined on its domain U^z. In general, a fuzzy set f_z (of a given variable z) is formally specified as $f_z = \{(x, \mu_{f_z}(x)) | x \in U^z, \mu_{f_z}(x) \in [0,1]\}$, where $\mu_{f_z}(x) \in [0,1]$ is the membership function of f_z. In the current research, considering that any link property is of a non-negative value (see later), for simplicity and computational efficiency, each fuzzy set is represented with a triangular membership function, except for certain properties that may involve an indefinite upper bound. A triangular membership function is specified by a tuple (x_1, x_2, x_3):

$$\mu_{f_z}(x) = \begin{cases} 0, & x < x_1 \\ \frac{x-x_1}{x_2-x_1}, & x_1 \leq x \leq x_2 \\ \frac{x_3-x}{x_3-x_2}, & x_2 \leq x \leq x_3 \\ 0, & x > x_3 \end{cases} \quad (7)$$

where x_1, x_3 are the left and right bounds, respectively, x_2 is the mode of the fuzzy set f_z (i.e. $\mu_{f_z}(x_2) = 1$) and $x, x_1, x_2, x_3 \in U^z$. For an indefinite boundary case, the membership function is defined as follows:

$$\mu_{f_z}(x) = \begin{cases} 0, & x < x_1 \\ \frac{x-x_1}{x_2-x_1}, & x_1 \leq x \leq x_2 \\ 1, & x \geq x_2 \end{cases} \quad (8)$$

Following the example of AOM model involving variables V_{CT} and V_{UQ}, V_{CT} can be expressed using the label set $L^{CT} = \{Small, Medium, Large\}$, whose quantitative semantics are represented by the collection of fuzzy sets F^{CT} shown in Fig 2(b). Similarly, the variable V_{UQ} is qualitatively described using the label set $L^{UQ} = \{Very\,Low, Low, Moderate, High, Very\,High\}$. Fig 3(b) presents the triangular fuzzy sets F^{UQ} that entail the semantics of these labels.

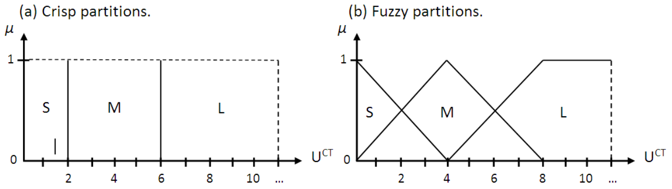

Figure 2: Order-of-magnitude partitions of the qualitative variable V_{CT} in: (a) original AOM model and (b) fuzzy AOM., where S = Small, M = Medium and L = Large.

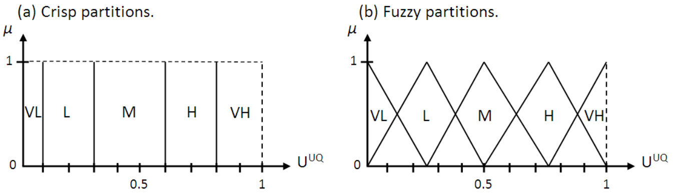

Figure 3: Order-of-magnitude partitions of the qualitative variable V_{UQ} in: (a) original AOM model and (b) fuzzy AOM, where VL = Very Low, L = Low, M = Moderate, H = High and VH = Very High.

With this representation scheme, an order-of-magnitude (OM) space S^a of a qualitative variable V_a is the combination of the ordered label set L^a and the mathematical definition of the collection F^a of the corresponding fuzzy sets. For example, the value of V_a may be expressed by the set of basic labels $L^a = \{B_1, \ldots, B_{n^a}\}$ and fuzzy definitions $F^a = \{\Pi_1, \ldots, \Pi_{n^a}\}$, where $B_1 < \ldots < B_{n^a}$ denotes the qualitative order amongst the fuzzy labels. The OM space S^a can be formally described as $S^a = L^a \cup \{[B_i(\mu_{\Pi_i}), B_j(\mu_{\Pi_j})] | B_i, B_j \in L^a, \Pi_i, \Pi_j \in F^a, i \leq j\}$. That is, the membership measures $\mu_{\Pi_i}, \mu_{\Pi_j} \in [0,1]$ specify the degree to which the underlying value can be expressed using labels B_i and B_j, respectively. This can be extended to support different forms of input value x_a (of variable V_a), such that

- When x_a is presented as a single label $B_i \in L^a$ and a union of labels $B_i, B_{i+1}, ..., B_{i+j} \in L^a$, it can be expressed as $[B_i, B_i]$ and $[B_i, B_j]$, respectively. Note that the membership degrees are irrelevant, and hence, they are omitted here. For instance, the value of variable V_{UQ} may be given as $[Moderate, Moderate]$ or $[Moderate, High]$.
- When x_a is presented as a real value (i.e. $x_a \in U^a$), it can be expressed as $[B_i(\mu_{\Pi_i(x_a)}), B_j(\mu_{\Pi_j(x_a)})]$, where $\mu_{\Pi_i(x_a)}$ and $\mu_{\Pi_j(x_a)}$ denotes the membership values of x_a to label-specific fuzzy sets $\Pi_i, \Pi_j \in F^a$. A quantity

of $V_{UQ} = 0.6$ that just falls into the 'Moderate' category in the interval-based AOM model, may actually be perceived and expressed differently from one analyst to another. With the fuzzy approach, this variation is generally captured through degrees that this figure belongs to distinct value sets. As a result, the given measure may be regarded as 'Moderate' or 'High', but with different membership degrees of 0.6 and 0.4, respectively (i.e. $[Moderate(0.6), High(0.4)]$).

- When x_a is presented as an interval of real values (i.e. $x_a = [x_{a1}, x_{a2}], x_{a1}, x_{a2} \in U^a, x_{a1} < x_{a2}$), it can be expressed by $[B_i(\mu_{\Pi_i(x_{a1})}), B_j(\mu_{\Pi_j(x_{a1})})]$ and $[B_p(\mu_{\Pi_p(x_{a2})}), B_q(\mu_{\Pi_q(x_{a2})})]$ that denote the lower and upper bound of x_a (where $B_i, B_j, B_p, B_q \in L^a$ and $\Pi_i, \Pi_j, \Pi_p, \Pi_q \in F^a$). As such, the measure of $V_{UQ} = [0.4, 0.6]$ is expressed by the composition of: $\{[Low(0.4),\ Moderate(0.6)],\ [Moderate(0.6),\ High(0.4)]\}$, where the former corresponds to the lower bound and the latter represents the upper bound.

3.2 Homogenisation of Multiple-Granularity Domains

The development of a qualitative reasoner usually involves a number of variables that are represented with qualitative labels of different granularity, defined on dissimilar universe of discourses. Therefore, the homogenisation process conducted in the conventional AOM model is similarly required to map the values of fuzzy variables onto a unified scale. This homogenisation can be summarised as follows:

- *Step 1*: For a variable V_D of an indefinite upper bound, whose discourse $U^D = [\alpha, \infty], \alpha \geq 0$, truncate the underlying discourse such that $U^D = [\alpha, \delta], \alpha < \delta < \infty$, where δ is the smallest element which has the full membership in the boundary case. Equivallently, δ now represents any real value $t \in [\delta, \infty)$. Note that as a result, this truncation helps to minimise the potential impact of certain unduely high values to dominate over the other possible values.
- *Step 2*: Normalise the discourse U^D of each variable V_D such that it is mapped onto the unified discourse $U^* = [0, 1]$. Principally, each value $g \in U^D$ is mapped to its corresponding value $g^* \in U^*$ by
$$g^* = \frac{g - \alpha}{\beta - \alpha} \tag{9}$$
where $\alpha, \beta \in \mathcal{R}$ denote the lower and upper bound of the discourse U^D (i.e. $U^D = [\alpha, \beta]$), respectively. $\beta = \delta$ for indefinite upper bound cases, of course.
- *Step 3*: Linearly map the fuzzy sets $F^D = \{f_1^D, \ldots, f_{n^D}^D\}$ of a variable V_D onto the discourse U^D onto the unified discourse U^* such that
$$\mu_{f_j^{D*}}(g^*) = \mu_{f_j^D}(g),\ \forall j = 1 \ldots n^D \tag{10}$$
where f_j^{D*}, which is specified on the unified discourse U^*, is the equivalent fuzzy set of f_j^D defined on the original U^D.

Given this procedure, the label sets of different granularity can be effectively homogenised. Particularly to the current example with variables V_{CT} and V_{UQ}, mapping the universe of discourses U^{UQ} to U^* is straight forward as it has already been defined on the $[0, 1]$ interval. Homogenising U^{CT} onto the common scale U^* is done via firstly truncating the range of discourse U^{CT} from $[0, \infty]$ to $[0, 8]$, such that the real value '8' represents any value in the interval $[8, \infty)$. Then each value $x \in U^{CT}$ is mapped onto its corresponding value $x^* \in U^*$ such that $x^* = \frac{x}{8}$ and $\mu_{f_j^{CT*}}(x^*) = \mu_{f_j^{CT}}(x), \forall j = 1 \ldots n^{CT}$.

3.3 Aggregation of Qualitative Descriptors

Many data analysis systems achieve a conclusion or final measure by aggregating values of different domain attributes. In general, each examined variable V_a can be assigned with a different degree of relevance (weight) W_a. This may be given by domain experts or estimated from past data if such knowledge is not readily available. Using the original AOM methodology, a weight can be expressed using the order-of-magnitude label set L^W, such as $L^W = \{None, +, ++, +++\}$ or $L^W = \{0, 1, 2, 3\}$. However, these crisp-interval descriptors are simply ineffective. The specification and subsequent manipulation of weights are more efficiently handled using the fuzzy approach.

For a variable V_a, its weight W_a, which is measured in the universe of discourse $U^W = [0, 1]$, is described using the label set $L^W = \{l_1^W, \ldots, l_{n^W}^W\}$, where n^W is the preferred number of qualitative labels. The semantics of each label $l_t^W, t = 1 \ldots n^W$ is captured by the corresponding fuzzy set $f_t^W \in F^W$. For instance, regarding the variable V_{UQ} in the previous example, the underlying label set can be represented as $L^W = \{Very\ Low, Low, Moderate, High, Very\ High\}$, where the corresponding fuzzy sets $f_t^W, t = 1 \ldots 5$ are similar to those defined in Fig 3(b).

With a problem involving m qualitative variables $(V_1, \ldots, V_m)$, their aggregated outcome Ω that is also a fuzzy set in U^*, can be estimated as follows:

$$\Omega = \varphi(V_1, \ldots, V_m, W_1, \ldots, W_m) = \frac{(V_1 W_1 + \ldots + V_m W_m)}{W_1 + \ldots + W_m} \tag{11}$$

where V_g and $W_g, g = 1 \ldots m$ are the qualitative descriptors of these variables and their weights, specified as fuzzy sets on the common universe of discourse $U^* = [0, 1]$. The membership function of Ω is denoted by $\mu_\Omega(t), \forall t \in U^*$, where t is an ordinary weighted average that is calculated as

$$t = \varphi(x_1, \ldots, x_m, w_1, \ldots, w_m) = \frac{x_1 w_1 + \ldots + x_m w_m}{w_1 + \ldots + w_m} \tag{12}$$

where $x_g \in V_g$ and $w_g \in W_g, g = 1 \ldots m$. By the extension principle, the membership function of Ω is:

$$\mu_\Omega(t) = \sup\Big(\min(\mu_{V_1}(x_1), \mu_{W_1}(w_1)), \ldots, \min(\mu_{V_m}(x_m), \mu_{W_m}(w_m))\Big) \tag{13}$$

Thus, finding the exact membership function $\mu_\Omega(t)$ is complicated and computationally expensive. Recognizing this, a discrete approximate method that makes use of the α-cut fuzzy arithmetic, is exploited to aggregate fuzzy sets (see [Chang *et al.*, 2006] for more details). In particular, the α-cut of a variable V_g and its weight $W_g, g = 1 \dots m$ are

$$(V_g)_\alpha = \{(x_g, \mu_{V_g}(x_g)) | x_g \in V_g, \mu_{V_g}(x_g) \geq \alpha\} \quad (14)$$

$$(W_g)_\alpha = \{(w_g, \mu_{W_g}(w_g)) | w_g \in W_g, \mu_{W_g}(w_g) \geq \alpha\} \quad (15)$$

These α-cuts are crisp intervals and can be expressed in continuous closed form as

$$(V_g)_\alpha = [(a_g)_\alpha, (b_g)_\alpha] = \Big[\min\{x_g \in V_g | \mu_{V_g}(x_g) \geq \alpha\}, \max\{x_g \in V_g | \mu_{V_g}(x_g) \geq \alpha\}\Big] \quad (16)$$

$$(W_g)_\alpha = [(c_g)_\alpha, (d_g)_\alpha] = \Big[\min\{w_g \in W_g | \mu_{W_g}(w_g) \geq \alpha\}, \max\{w_g \in W_g | \mu_{W_g}(w_g) \geq \alpha\}\Big] \quad (17)$$

where $(a_g)_\alpha$ and $(c_g)_\alpha$ are the *left endpoints* of $(V_g)_\alpha$ and $(W_g)_\alpha, g = 1 \dots m$, respectively. Analogously, $(b_g)_\alpha$ and $(d_g)_\alpha$ are the *right endpoints* of $(V_g)_\alpha$ and $(W_g)_\alpha, g = 1 \dots m$.

From this, the α-cut $(\Omega)_\alpha$ can be obtained such that

$$(\Omega)_\alpha = \Big[\min \varphi(x_1, \dots, x_m, w_1, \dots, w_m), \max \varphi(x_1, \dots, x_m, w_1, \dots, w_m)\Big] \quad (18)$$

where $\forall \alpha \in (0, 1]$, $(a_g)_\alpha \leq x_g \leq (b_g)_\alpha$, $(c_g)_\alpha \leq w_g \leq (d_g)_\alpha$, $a_g, x_g, b_g \in V_g$ and $c_g, w_g, d_g \in W_g, g = 1 \dots m$. Due to the monotonicity of the function φ, this equation can be simplified to become

$$(\Omega)_\alpha = \Big[\min_{w_g \in \{(c_g)_\alpha, (d_g)_\alpha\}} f_L(w_1, \dots, w_m), \max_{w_g \in \{(c_g)_\alpha, (d_g)_\alpha\}} f_R(w_1, \dots, w_m)\Big] \quad (19)$$

where $f_L(w_1, \dots, w_m)$ and $f_R(w_1, \dots, w_m)$ are defined as

$$f_L(w_1, \dots, w_m) = f((a_1)_\alpha, \dots, (a_m)_\alpha, w_1, \dots, w_m) = \frac{(a_1)_\alpha w_1 + \dots + (a_m)_\alpha w_m}{w_1 + \dots + w_m} \quad (20)$$

$$f_R(w_1, \dots, w_m) = f((b_1)_\alpha, \dots, (b_m)_\alpha, w_1, \dots, w_m) = \frac{(b_1)_\alpha w_1 + \dots + (b_m)_\alpha w_m}{w_1 + \dots + w_m} \quad (21)$$

4 Qualitative Model of Link Analysis

4.1 Problem Formulation

Link analysis can be conducted on a social network representation of relations amongst references of real-world entities. This has been effectively exploited for link prediction in [Liben-Nowell and Kleinberg, 2007], [Murata and Moriyasu, 2008] and author-name resolution in [Reuther and Walter, 2006]. An underlying link network is specified as an undirected graph $G = (V, E)$, which is composed of two sets: the set of vertices V and that of edges E. Let X and R be the set of all references and that of their relations in the dataset, respectively. Then, vertex $v_i \in V$ denotes reference $x_i \in X$ and each edge $e_{ij} \in E$ linking vertices $v_i \in V$ and $v_j \in V$ corresponds to a relation $r_{ij} \in R$ between the references $x_i \in X$ and $x_j \in X$.

The current research concentrates on analysing a social network whose edges correspond to 'co-occurrence' relations amongst references. A relation $r_{ij} \in R$ determines the fact that references $x_i, x_j \in X$ appear together in a specific observation. It is bi-directional such that r_{ij} is equivalent to $r_{ji}, \forall r_{ij}, r_{ji} \in R$. As a result, edges in G are undirected and any $e_{ij}, e_{ji} \in E$ are equivalent. This approach is simple and efficient regarding its associated procedures for information acquisition and analysis.

Each edge $e_{ij} \in E$ possess statistical information $f_{ij} \in \{1, \dots, \infty\}$, reflecting the frequency of a relation between references x_i and x_j within the underlying dataset. In so doing, the graph terminology used here becomes simple (i.e. with no parallel edges), without losing any potential link information [Wasserman and Faust, 1994]. Let O be the set of real-world entities, each being referred to by at least one member of set X, a pair of references (x_i, x_j) are aliases 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 greater the possibility of vertices v_i and v_j, and hence the more likely that the corresponding references x_i and x_j constitute the actual alias pair.

4.2 Properties of Link Patterns

Link analysis makes use of the link-based similarity that is measured upon link patterns between studied references. A number of link-based similarity methods exist that evaluate the similarity between information objects, including: SimRank [Jeh and Widom, 2002], PageSim [Lin *et al.*, 2006] and a variety of random walk methods [Fouss *et al.*, 2007], [Minkov *et al.*, 2006] (see more details in [Getoor and Diehl, 2005], [Liben-Nowell and Kleinberg, 2007]). Despite notable achievement, these techniques are computationally inefficient compared to those simple methods developed for social network analysis (SNA) [Wasserman and Faust, 1994]. To further boost the performance of the SNA methodology whilst maintaining its efficiency, the fuzzy aggregation model is employed to provide a systematic framework for combining multiple link properties and consequently deriving an accurate conclusion. Specifically, the following property measures are employed in the present research to illustrate the potential of this approach.

- *Cardinality (CT)*: The similarity amongst social members can be decided upon their 'common neighbours'. With a social network being represented as an undirected graph $G = (V, E)$, neighbours $N_{v_i} \subset V$ of each $v_i \in V$ form a set of vertices directly linked to v_i, i.e.

$e_{ij} \in E, v_i \neq v_j, \forall v_i, v_j \in V$. Effectively, the common neighbours of $v_i, v_j \in V$ can be identified as $N_{v_i} \cap N_{v_j}$. The similarity between v_i and v_j is then determined by the 'cardinality' of their shared neighbours, $|N_{v_i} \cap N_{v_j}|$.

- *Uniqueness (UQ)*: Cardinality based methods are simple, but greatly sensitive to noise, often generating a large proportion of false positives [Reuther and Walter, 2006]. To refine the estimation of similarity values, the 'uniqueness' measure has been suggested as the additional criterion to CT [Boongoen *et al.*, 2010]. Given a graph $G(V, E)$, a uniqueness measure UQ_{ij}^k of any two objects i and j (denoted by vertices $v_i, v_j \in V$) can be approximated from each joint neighbour k (denoted by the vertex $v_k \in V$) as follows:

$$UQ_{ij}^k = \frac{f_{ik} + f_{jk}}{\sum_m f_{mk}} \tag{22}$$

where f_{ik} is the frequency of the link between objects i and k occurring in data, f_{jk} is the frequency of the link between objects j and k, and f_{mk} is the frequency of the link between object k and any object m. To summarise the uniqueness of joint link patterns UQ_{ij} between objects i and j, the ratios estimated for each shared neighbour are aggregated as

$$UQ_{ij} = \frac{1}{n} \sum_{k=1}^{n} UQ_{ij}^k \tag{23}$$

where n is the number of overlapping neighbour objects that objects i and j are commonly linked to.

4.3 Aggregation-Based Similarity Evaluation

The semantics of numerical similarity and link-property values are subject to individual perception and can become inefficient for coherent human interpretation and communication. Besides, numerical descriptions of a property measure may be inaccurate, usually due to a few link patterns with unduly high values. Such skewed distribution of link patterns often causes the others to be overlooked. The case may become worse when such a property is aggregated with other properties. The fuzzy AOM approach helps to make the process of similarity estimation more accurate and interpretable.

A fuzzy AOM model is herein established with qualitative variables V_{CT} and V_{UQ}, which stand for the aforementioned CT and UQ property measures. Suppose that the corresponding label sets are $L^{CT} = \{Small, Medium, Large\}$ and $L^{UQ} = \{Very\ Low, Low, Moderate, High, Very\ High\}$. These qualitative descriptors are semantically defined by the collections of fuzzy sets F^{CT1} and F^{UQ1}, matching those presented in Fig 2(b) and Fig 3(b), respectively. To generalise the performance of the proposed aggregation model, additional partitions of fuzzy sets (F^{CT2} and F^{UQ2}, see Fig 4) that define label sets of L^{CT} and L^{UQ}, are also employed in this study. The weights W_{CT} and W_{UQ} (of V_{CT} and V_{UQ}) are expressed by the label set L^W and fuzzy sets F^W, which are identical to L^{UQ} and F^{UQ1}.

Based on Equations 19-21, the fuzzy OMS that represents the order-of-magnitude similarity, can be estimated from property measures (V_{CT} and V_{UQ}) and their weights W_{CT} and W_{UQ} such that

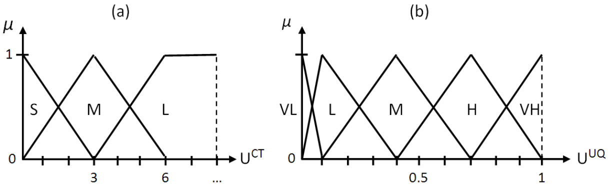

Figure 4: Definition of fuzzy sets for (a) F^{CT2} and (b) F^{UQ2}, used in the fuzzy AOM model.

$$(OMS)_\alpha = \Big[\min_{w_g \in \{(c_g)_\alpha, (d_g)_\alpha\}} f_L(w_{ct}, w_{uq}),$$
$$\max_{w_g \in \{(c_g)_\alpha, (d_g)_\alpha\}} f_R(w_{ct}, w_{uq}) \Big] \tag{24}$$

where $w_{ct}, c_{ct}, d_{ct} \in W_{CT}$, $w_{uq}, c_{uq}, d_{uq} \in W_{UQ}$, $f_L(w_{ct}, w_{uq})$ and $f_R(w_{ct}, w_{uq})$ are defined by

$$f_L(w_{ct}, w_{uq}) = f((a_{ct})_\alpha, (a_{uq})_\alpha, w_{ct}, w_{uq}) = \frac{(a_{ct})_\alpha w_{ct} + (a_{uq})_\alpha w_{uq}}{w_{ct} + w_{uq}} \tag{25}$$

$$f_R(w_{ct}, w_{uq}) = f((b_{ct})_\alpha, (b_{uq})_\alpha, w_{ct}, w_{uq}) = \frac{(b_{ct})_\alpha w_{ct} + (b_{uq})_\alpha w_{uq}}{w_{ct} + w_{uq}} \tag{26}$$

where $a_{ct}, b_{ct} \in V_{CT}$ and $a_{uq}, b_{uq} \in V_{UQ}$.

To support consistent interpretation and comparison, OMS values are mapped onto the standard ordered set $L^S = \{l_1^S, \ldots, l_{n^S}^S\}$ of n^S qualitative labels, specified on the universe of discourse $U^* = [0, 1]$. The semantics of each label $l_t^S, t = 1 \ldots n^S$ is represented by the fuzzy set $f_t^S \in F^S$, $F^S = \{f_1^S, \ldots, f_{n^S}^S\}$. As an example, fuzzy OMS can be expressed using $L^S = \{Very\ Low, Low, Moderlate, High, Very\ High\}$, where the corresponding fuzzy sets $f_t^S, t = 1 \ldots 5$ are identical to those of F^{UQ1} (see Fig 3(b)). Given this, OMS is mapped onto each label in L^S as follows:

$$OMS = (l_j^S, \beta_j^S(OMS)), j = 1 \ldots n^S \tag{27}$$

where $\beta_j^S(OMS)$ is defined by

$$\beta_j^S(OMS) = \max_{\forall t \in U^*} \min(\mu_{f_j^S}(t), \mu_{OMS}(t)) \tag{28}$$

5 Application of Qualitative Link Analysis to Alias Detection in Intelligence Data

5.1 False Identity Detection in Intelligence Data

Identity is a set of characteristic descriptors unique to a specific person, which can be principally categorised into three types of attributed, biographical and biometric identity, respectively [Clarke, 1994], [Wang *et al.*, 2006]. Initially with attributed identity, a person can be identified using descriptions of name, details of parents, date and place of birth. Biographical identity constituted from personal information over a life span (e.g. criminal, educational and financial history) can also be exploited for such purpose. Comparing to biometric identity like fingerprints and DNA features, the first two types are greatly subject to deception as they are much easier to falsify. 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. Note that name deception is the practice commonly adopted by criminal cases [Wang *et al.*, 2006].

To appreciate the challenge facing false identity detection, the 'Terrorist' data 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 the name of a person (criminal/terrorist), place or organisation, while each of the $5,581$ links denotes the co-occurrence of a specific pair of names. Existing models developed with regard to this dataset are dissimilar to those of using the OMS and other link-based measures, due to their fundamental differences in the learning schemes adopted (i.e. supervised vs. unsupervised).

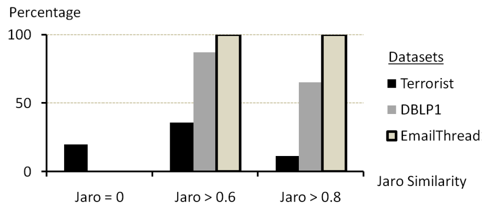

Figure 5: Percentage of true alias pairs in different datasets, categorised in accordance with their text-based similarity (using Jaro).

Figure 5 shows the number of alias pairs in Terrorist and other data collections (see [Shen and Boongoen, 2010] for details about these datasets), with respect to different settings of the Jaro string-matching measures [Navarro, 2001]. Clearly, name ambiguity in intelligence data such as Terrorist is highly subject to intentional deception. This differs from the problem regarding publication data (DBLP1), which is caused mostly by human entry errors, and the problem regarding email data (EmailThread1), which originates from an automated identification system. Typical string-matching techniques [Navarro, 2001] are effective to discover only a small number of aliases, whose matching scores are very high. They fail to reveal the association between the following pairs of terrorists' names, whose overlapping text content is void: ('ashraf refaat nabith henin', 'salem ali'), ('bin laden', 'the prince'), ('bin laden', 'the emir') and ('abu mohammed nur al-deen', 'the doctor'). It is therefore interesting to investigate if fuzzy OMS can tackle the burden of disclosing such unconventional truth.

5.2 Performance Evaluation of OMS Measure

In this work, modelling parameters such as fuzzy-set definitions in describing the OMS values, link-property measures and corresponding weights are subjectively designed by domain experts, and hence are data-independent. Such parameter settings are not assumed to have been optimised and their use may lead to performance variations with respect to different datasets that exhibit different characteristics.

To better gauge the effectiveness of the proposed approach without resorting to substantial overheads of computation, two weighting schemes ($W1 = \{W_{CT} = High, W_{UQ} = Moderate\}$ and $W2 = \{W_{CT} = High, W_{UQ} = High\}$) and two different collections of property-specific fuzzy sets ($\{F^{CT1}, F^{UQ1}\}$ and $\{F^{CT2}, F^{UQ2}\}$) are employed (see details in subsection 4.3). This results in the following OMS measures, which are investigated herein: OMS$_1$ $\leftarrow \{F^{CT1}, F^{UQ1}, W1\}$, OMS$_2$ $\leftarrow \{F^{CT1}, F^{UQ1}, W2\}$, OMS$_3$ $\leftarrow \{F^{CT2}, F^{UQ2}, W1\}$ and OMS$_4$ $\leftarrow \{F^{CT2}, F^{UQ2}, W2\}$. These OMS measures are assessed against various unsupervised methods that have been developed for a similar problem, including:

- Basic property measures that are exploited to construct the proposed OMS methods: numerical CT and UQ.
- Well-known link-based similarity techniques: SimRank (SR) [Jeh and Widom, 2002] and PageSim (PS) [Lin *et al.*, 2006]. SimRank was introduced for evaluating similarity amongst objects in citation and web-page graphs. Its underlying mechanism relies on the normalised CT measure, which is refined through an iterative expansion of neighbouring context. Likewise, PageSim was developed to capture similar web pages based on associations implied by their hyperlinks. To determine the proximity between web pages, it exploits the random-walk mechanism over a link network, in which nodes (corresponding to web pages) are ranked using PageRank [Brin and Page, 1998] of the Google search engine.
- String-matching techniques: Jaro (JR), Levenshtein (LT), Q-grams (QG) and Needleman-Wunsch (NW) (see details in [Navarro, 2001]). To consolidate the investigation with this paradigm, the aggregation of different string-matching scores [Branting, 2002], [Lyras *et al.*, 2008] is also examined. For matching scores $s_\gamma(st_1, st_2) \in [0,1]$ between string st_1 and st_2 that are estimated by the algorithm $\gamma \in$ {JR, LT, QG, NW}, its aggregated measure $s^*(st_1, st_2)$ is

$$s^*(st_1, st_2) = \sum_{\forall \gamma \in \{JR, LT, QG, NW\}} w_\gamma s_\gamma(st_1, st_2) \quad (29)$$

  where $w_\gamma \in [0,1]$ and $\sum_{\forall\gamma} w_\gamma = 1$. The simplest aggregation model (denoted by TXT-AVG) exploits the arithmetic average, such that $w_\gamma = \frac{1}{n}$, where n denotes the number of matching scores to be aggregated. For the current configuration with n being 4, each $w_\gamma = 0.25$. Another aggregation model (TXT-REL) is also investigated here. It implements the weight determination procedure of [Boongoen and Shen, 2010], where a higher weight is allocated to more reliable input-argument. The reliability of each score is determined by the distance to its k-nearest neighbours ($k = 1$ in this assessment).
- Computational linguistic methods: BN [Baroni *et al.*, 2002] and SC [Schone and Jurafsky, 2000] which discover morphologically related words in an unannotated

corpus. These techniques work without employing *a priori* knowledge regarding to the linguistic properties of investigated words. To justify the similarity $s(p_1, p_2)$ of any pair of words (p_1 and p_2), they combine the corresponding appearance-based and semantic-based similarity measures ($\Lambda(p_1, p_2) \in [0, 1]$ and $\Phi(p_1, p_2) \in [0, 1]$, respectively):

$$s(p_1, p_2) = w_\Lambda \Lambda(p_1, p_2) + w_\Phi \Phi(p_1, p_2) \qquad (30)$$

where $w_\Lambda, w_\Phi \in [0, 1]$ and $w_\Lambda + w_\Phi = 1$.

In particular, for the BN model, the semantic similarity of $\Phi(p_1, p_2)$ is determined by the frequency of co-occurrence between p_1 and p_2. For the SC method, this measure is defined by the correlation between vectors T_{p_1} and T_{p_2}, containing the frequencies (presented as z-scores) in which p_1 and p_2 co-occur with other words in the dataset. The semantic similarity is estimated from the cosine of the angle between T_{p_1} and T_{p_2}:

$$\cos(T_{p_1}, T_{p_2}) = \frac{T_{p_1} \, . \, T_{p_2}}{||T_{p_1}|| \, ||T_{p_2}||} \qquad (31)$$

In conducting the performance evaluation, JR and LT are used to estimate the two appearance-based similarities, respectively. The final result is achieved using a simple arithmetic average (i.e. $w_\Lambda = w_\Phi = \frac{1}{2}$). Given this configuration, semantic similarity (BN and SC) and combined measures (BN-JR, BN-LT, SC-JR and SC-LT) can be obtained.

Table 2 compares the numbers of successfully disclosed alias pairs by the examined methods over the Terrorist data, with respect to the number of retrieved entity pairs (K) of the highest similarity measures. These results suggest that the OMS methods substantially improve the effectiveness of the elementary measures, CT and UQ. In addition, they usually outperform the more complex SR and PS, and the text-based techniques investigated. Note that the precision and recall measures are estimated by

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

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

Table 2: Number of alias pairs disclosed by each method, with the corresponding precision/recall given in brackets.

Method	**K=400**	**K=600**	**K=800**
OMS Measures			
OMS$_1$	54 (0.135/0.059)	85 (0.142/0.092)	129 (0.161/0.140)
OMS$_2$	57 (0.143/0.062)	77 (0.128/0.084)	112 (0.140/0.122)
OMS$_3$	116 (0.290/0.126)	148 (0.246/0.161)	151 (0.189/0.164)
OMS$_4$	104 (0.260/0.113)	145 (0.242/0.158)	160 (0.200/0.174)
Link-Based			
CT	5 (0.012/0.005)	5 (0.008/0.005)	77 (0.096/0.084)
UQ	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)
SR	0 (0.0/0.0)	1 (0.002/0.001)	1 (0.001/0.001)
PS	36 (0.090/0.039)	63 (0.105/0.069)	79 (0.098/0.086)
String-Matching			
JR	33 (0.083/0.036)	40 (0.067/0.044)	43 (0.054/0.047)
LT	34 (0.085/0.037)	44 (0.073/0.048)	50 (0.063/0.054)
QG	31 (0.078/0.034)	37 (0.062/0.040)	46 (0.058/0.050)
NW	36 (0.090/0.039)	41 (0.068/0.045)	48 (0.060/0.052)
TXT-AVG	35 (0.088/0.038)	42 (0.070/0.046)	47 (0.059/0.051)
TXT-REL	36 (0.090/0.039)	45 (0.075/0.049)	52 (0.065/0.057)
Morpheme-Based			
BN	1 (0.003/0.001)	0 (0.0/0.0)	0 (0.0/0.0)
BN-JR	0 (0.0/0.0)	0 (0.0/0.0)	1 (0.001/0.001)
BN-LT	0 (0.0/0.0)	1 (0.002/0.001)	1 (0.001/0.001)
SC	0 (0.0/0.0)	0 (0.0/0.0)	1 (0.001/0.001)
SC-JR	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)
SC-LT	0 (0.0/0.0)	0 (0.0/0.0)	0 (0.0/0.0)

To illustrate the deficiency of the numerical link-based methods, Table 3 presents the number of retrieved entity pairs (K), each with a similarity value greater than a given threshold (α), and the corresponding number (D) of disclosed alias pairs. These results show that both PS and CT generate inaccurate similarity descriptions. This is because certain entity pairs possess unduly high measures, which mislead the interpretation of those belonging to other pairs. Even though SR and UQ do not encounter this problem, they are ineffective for alias detection giving low precision statistics. Furthermore, based on Table 4, qualitative descriptions (e.g. 'High') generated by the OMS measures are more accurate than those by CT and PS. This indicates that the OMS approach is effective to tackle the problem of exceptionally high-valued cases within a network. Additionally, the precisions of these qualitative methods are substantially greater than those of their SR and UQ counterparts.

Table 3 also includes the result of TXT-RE (threshold-based performance analysis), a representative text-based approach. It is remarkably effective to discover the minority of alias pairs with very high appearance similarity (e.g. when similarity > 0.9). However, as compared to the OMS measures, its performance decreases drastically with slightly lower threshold values (e.g. 0.8 or 0.7). It even becomes immaterial with the 'highly deceptive case' where overlapping textual content is void (i.e. TXT-REL similarity is 0).

For the 183 name pairs in Terrorist that are highly deceptive, Fig 6 presents the number of such pairs that are revealed by different link-based methods (regarding the number of those retrieved pairs (K) of the highest similarity values). This result indicates that the OMS approach is effective for tackling the problem of deception detection, with the performance being generally robust to different parameter settings.

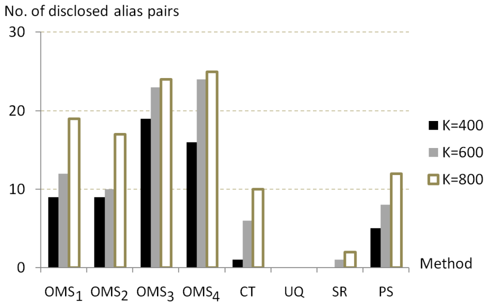

Figure 6: The number of 'highly deceptive' alias pairs discovered from K name pairs of the highest similarity measures.

Table 3: Numbers of retrieved (K) and discovered (D) alias pairs by examined methods at different thresholds (α), where the corresponding (precision/recall) measures are given in brackets.

α	**PS**		**SR**		**CT**		**UQ**		**TXT-REL**	
	K	D	K	D	K	D	K	D	K	D
0.9	3	0 (0.0/0.0)	1,587	2 (0.001/0.002)	3	0 (0.0/0.0)	497	0 (0.0/0.0)	343	31 (0.090/0.034)
0.8	3	0 (0.0/0.0)	1,597	2 (0.001/0.002)	3	0 (0.0/0.0)	512	0 (0.0/0.0)	2,251	104 (0.046/0.113)
0.7	3	0 (0.0/0.0)	1,657	2 (0.001/0.002)	3	0 (0.0/0.0)	558	0 (0.0/0.0)	23,178	209 (0.009/0.227)
0.6	3	0 (0.0/0.0)	2,272	2 (0.0008/0.002)	3	0 (0.0/0.0)	1,245	0 (0.0/0.0)	377,056	328 (0.001/0.357)
0.5	3	0 (0.0/0.0)	4,582	2 (0.0004/0.002)	3	0 (0.0/0.0)	1,938	1 (0.0005/0.001)	2,393,002	513 (0.000/0.558)
0.4	4	0 (0.0/0.0)	5,755	28 (0.005/0.030)	3	0 (0.0/0.0)	2,625	6 (0.002/0.006)	5,122,131	709 (0.000/0.771)
0.3	5	0 (0.0/0.0)	10,224	125 (0.012/0.136)	3	0 (0.0/0.0)	4,041	51 (0.012/0.055)	5,743,556	732 (0.000/0.797)
0.2	14	0 (0.0/0.0)	18,064	175 (0.009/0.190)	11	0 (0.0/0.0)	8,441	237 (0.028/0.258)	5,811,998	736 (0.000/0.801)
0.1	61	0 (0.0/0.0)	36,743	271 (0.007/0.295)	67	0 (0.0/0.0)	18,642	328 (0.017/0.357)	5,811,998	736 (0.000/0.801)
0.0	708,613	468 (0.0006/0.51)	708,613	468 (0.0006/0.51)	81,985	366 (0.004/0.398)	81,985	366 (0.004/0.398)	5,811,998	736 (0.000/0.801)

Table 4: Numbers of retrieved (K) and discovered (D) alias pairs by the OMS measures with different thresholds (α), where the corresponding precision/recall measures are given in brackets.

α	$\mathbf{OMS_1}$		$\mathbf{OMS_2}$		$\mathbf{OMS_3}$		$\mathbf{OMS_4}$	
	K	D	K	D	K	D	K	D
$\geq VH$	76	3 (0.040/0.003)	76	3 (0.040/0.003)	32	3 (0.094/0.003)	28	2 (0.071/0.002)
$\geq H$	1,566	188 (0.120/0.205)	1,566	188 (0.120/0.205)	2,304	241 (0.105/0.262)	2,290	237 (0.103/0.258)
$\geq M$	12,814	251 (0.020/0.273)	12,814	251 (0.020/0.273)	14,253	258 (0.018/0.281)	14,281	258 (0.018/0.281)
$\geq L$	81,985	366 (0.004/0.398)	81,985	366 (0.004/0.398)	81,985	366 (0.004/0.398)	81,985	366 (0.004/0.398)

5.3 Explanatory Support with OMS Approach

With the OMS approach, it is feasible to present the similarity assessment in terms of ordered linguistic labels, such as 'High' or 'Low'. This ability of computing-with-words supports coherent interpretation and natural communication amongst intelligence data analysts. This is hardly achievable through numerical measures whose values may often be inconsistently justified upon analysts' own evaluation scales, which may be deeply dictated by personal experience, bias and urgency of response. For the present research, qualitative descriptors of link property measures, their weights and the ultimate similarity degree are designed by human experts, each covering a range of possibilities (though usually to a certain degree). This is in line with the linguistic approach to social network analysis, recently established in [Yager, 2008].

With examples from the Terrorist data, Table 5 illustrates a variety of the OMS similarity classes and their supporting measures of CT and UQ. Note that the following sets of abbreviation are used for simplicity: (VH = Very High, H = High, M = Medium, L = Low, VL = Very Low, D = Dissimilar) for the OMS and UQ qualitative values, and (VL = Very Large, M = Medium, S = Small) for the CT measures. It can be seen that the similarity of an entity pair is not only categorised into a qualitative class over the coherently-interpretable scale, but the underlying causes of such justification can also be linguistically explained. For instance, the similarity of the entity pair ('abu abdallah', 'the teacher') is 'Very High(0.1), High(0.65)' because its corresponding cardinality and uniqueness measures of their joint link patterns are 'Large(0.25), Medium(0.75)' and 'Very High(1.0)', respectively.

Such an explanatory mechanism can assist data analysts to validate the results generated by the OMS approach. This capability can help in reducing the problem of false positives, which is usually difficult to address with quantitative link-based methods. In particular, an analyst can better analogically compare the generated outcome with human concepts of true aliases and false decisions, which have been built upon first-hand experience of resolving past cases. Also, the explanatory formalism allows a flexible, linguistic-like retrieval of suspected cases, which has proven effective for publications and online resources [Bordogna *et al.*, 2003], [Kraft *et al.*, 1998].

Clearly, the above experimental evaluations consistently demonstrate that the present work significantly outperforms typical existing approaches. Although the precision and recall rates achievable by this work may appear to be rather low for the Terrorist data, with regard to the usual expectation in other problem domains, it is remarkable to be able to reach such a level of detection. This is because the underlying true alias are typically intentionally deceptive and misleading. Any positive identification is of great importance in aiding intelligence data analysts, for instance, to draw their attention to such possible alias pairs which alternative approaches may fail to detect. Recent applications of this research in different problem domains such as student academic performance evaluation [Boongoen *et al.*, 2011] have indeed shown that the precision and recall rates can reach much higher values.

6 Conclusion

This paper has presented a novel approach to developing fuzzy set based order-of-magnitude model for qualitative link analysis. Unlike many link-based techniques that concentrate on one specific measure of common neighbours (e.g. cardinality), the proposed approach combines a number of link properties in order to refine the evaluation of similarity. As a result, this work allows coherent interpretation of, and reasoning with imprecise descriptions of numerical properties, which is hardly feasible with pure numerical terms. This explanatory entailment assists data analysts to validate the re-

Table 5: Examples of OMS similarity assessment, with membership degrees given in brackets.

Entity Pair	OMS	CT	UQ
ahmed majdalani-qiblani, abu iyad	VH(0.53), H(0.20)	L(1.00)	H(0.83), M(0.17)
abu abdallah, the teacher	VH(0.10), H(0.65)	L(0.25), M(0.75)	VH(1.00)
saif al-adel, sheik ahmed yassin	H(0.36), M(0.41)	L(0.75), M(0.25)	L(0.67), VL(0.33)
zaheer ul-islam abbasi, pervez musharraf	H(0.15), M(0.61)	M (1.00)	H(0.5), M(0.5)
osama bin ladin, osama bin laden	L(0.68), VL(0.07)	M(0.75), S(0.25)	L(0.64), VL(0.36)
abu ali mustafa, dr subhu ghosheh	L(0.27), VL(0.50)	M(0.25), S(0.75)	L(0.44), VL(0.56)

sults and to partially resolve the problem with false positives.

Empirically, this qualitative approach consistently outperforms typical existing methods over datasets available in the terrorism domain (as well as in the publication and email domains [Shen and Boongoen, 2010]). Although the order-of-magnitude model is currently parameterised by expert-directed specification of qualitative variables (link property measures, their weights and the resulting similarity), its performance is generally robust to different parameter settings. However, the underlying qualitative definitions may be automatically determined with a data-driven learning process, if relevant training data is available. This remains as active research.

References

[Adamic and Adar, 2003] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25:211–230, 2003.

[Agell *et al.*, 2000] N. Agell, X. Rovira, and C. Ansotegui. Homogenising references in orders of magnitude spaces: An application to credit risk prediction. In *Proceedings of 14th International Workshop on Qualitative Reasoning, Mexico*, pages 1–8, 2000.

[Ali *et al.*, 2003] A. H. Ali, D. Dubois, and H. Prade. Qualitative reasoning based on fuzzy relative orders of magnitude. *IEEE Transactions on Fuzzy Systems*, 11(1):9–23, 2003.

[Baeza and Ribeiro, 1999] Y. R. Baeza and N. B. Ribeiro. *Modern information retrieval.* Addison Wesley/ACM Press, 1999.

[Baroni *et al.*, 2002] M. Baroni, J. Matiasek, and H. Trost. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning*, pages 48–57, 2002.

[Bhattacharya and Getoor, 2007] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. *ACM Trans. on KDD*, 1(1), 2007.

[Bilenko and Mooney, 2003] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of ACM SIGKDD International Conference on KDD*, pages 39–48, 2003.

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

[Boongoen and Shen, 2010] T. Boongoen and Q. Shen. Nearest-neighbor guided evaluation of data reliability and its applications. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 40(6):1622–1633, 2010.

[Boongoen *et al.*, 2010] T. Boongoen, Q. Shen, and C. Price. A hybrid link analysis approach for false identity detection. *AI and Law*, 18(1):77–102, 2010.

[Boongoen *et al.*, 2011] T. Boongoen, Q. Shen, and C. Price. Fuzzy qualitative link analysis for academic performance evaluation. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 19(3):559–585, 2011.

[Bordogna *et al.*, 2003] G. Bordogna, G. Pasi, and R. R. Yager. Soft approaches to distributed information retrieval. *International Journal of Intelligent Systems*, 34:105–120, 2003.

[Branting, 2002] L. K. Branting. Name-matching algorithms for legal case-management systems. *Journal of Information, Law and Technology*, 2002(1), 2002.

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

[Chang *et al.*, 2006] P. T. Chang, K. C. Hung, K. P. Lin, and C. H. Chang. A comparison of discrete algorithms for fuzzy weighted average. *IEEE Transactions on Fuzzy Systems*, 14(5):663 – 675, 2006.

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

[Fellegi and Sunter, 1969] I. Fellegi and A. Sunter. Theory of record linkage. *Journal of the American Statistical Association*, 64:1183–1210, 1969.

[Fouss *et al.*, 2007] F. Fouss, A. Pirotte, J. M. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Trans. on Know. and Data Engineering*, 19(3):355–369, 2007.

[Getoor and Diehl, 2005] L. Getoor and C. P. Diehl. Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12, 2005.

[Hölzer *et al.*, 2005] R. Hölzer, B. Malin, and L. Sweeney. Email alias detection using social network analysis. In *Proceedings of International Workshop on Link Discovery*, pages 52–57, 2005.

[Hsiung *et al.*, 2005] P. Hsiung, A. Moore, D. Neill, and J. Schneider. Alias detection in link data sets. In *Proceedings of the International Conference on Intelligence Analysis*, 2005.

[Jeh and Widom, 2002] G. Jeh and J. Widom. Simrank: A measure of structural-context similarity. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 538–543, 2002.

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

[Kraft *et al.*, 1998] D. H. Kraft, G. Bordogna, and G. Pasi. Information retrieval systems: where is the fuzz? In *Proceedings of IEEE International Conference on Fuzzy Systems*, pages 1367–1372, 1998.

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

[Lin *et al.*, 2006] Z. Lin, I. King, and M. R. Lyu. Pagesim: A novel link-based similarity measure for the world wide web. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong*, pages 687–693, 2006.

[Lyras *et al.*, 2008] D. P. Lyras, K. N. Sgarbas, and N. D. Fakotakis. Applying similarity measures for automatic lemmalization: a case study for modern Greek and English. *International Journal on Artificial Intelligence Tools*, 17(5):1043–1064, 2008.

[Malin *et al.*, 2005] B. Malin, E. Airoldi, and K. M. Carley. A network analysis model for disambiguation of names in lists. *Computational and Mathematical Organization Theory*, 11:119–139, 2005.

[Minkov *et al.*, 2006] E. Minkov, W. W. Cohen, and A. Y. Ng. Contextual search and name disambiguation in email using graphs. In *Proceedings of 29th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34, 2006.

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

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

[Newman, 2003] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[Pantel, 2006] P. Pantel. Alias detection in malicious environments. In *Proceedings of AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection, Washington, D.C.*, pages 14–20, 2006.

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

[Popp and Yen, 2006] R. L. Popp and J. Yen. *Emergent information technologies and enabling policies for counterterrorism.* Wiley, 2006.

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

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

[Schone and Jurafsky, 2000] P. Schone and D. Jurafsky. Knowldedge-free induction of morphology using latent semantic analysis. In *Proceedings of the Conference on Computational Natural Language Learning*, 2000.

[Shen and Boongoen, 2010] Q. Shen and T. Boongoen. Fuzzy orders-of-magnitude based link analysis for qualitative alias detection. *IEEE Transactions on Knowledge and Data Engineering*, (DOI: 10.1109TKDE.2010.255), 2010.

[Shen and Leitch, 1992] Q. Shen and R. Leitch. On extending the quantity space in qualitative reasoning. *AI in Engineering*, 7:167–173, 1992.

[Sugeno and Yasukawa, 1993] M. Sugeno and T. Yasukawa. A fuzzy logic-based approach to qualitative modeling. *IEEE Transactions on Fuzzy Systems*, 1(1):7–31, 1993.

[Travé-Massuyès and Dague, 2003] L. Travé-Massuyès and P. Dague. *Modèles et raisonnements qualitatifs*. Lavoisier, Hermes Science, Paris, 2003.

[Travé-Massuyès and Piera, 1989] L. Travé-Massuyès and N. Piera. The orders of magnitude models as qualitative algebras. In *Proceedings of 11th International Joint Conference on Artificial Intelligence*, pages 1261–1266, 1989.

[Wang *et al.*, 2006] G. A. Wang, H. Chen, J. J. Xu, and H. Atabakhsh. Automatically detecting criminal identity deception: an adaptive detection algorithm. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 36(5):988–999, 2006.

[Wasserman and Faust, 1994] S. Wasserman and K. Faust. *Social network analysis: methods and applications*. Cambridge University Press, 1994.

[Yager, 2008] R. R. Yager. Intelligent social network analysis using granular computing. *Int. Jrn. of Intelligent Systems*, 23:1197–1220, 2008.