

Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection

Boanerges Aleman-Meza, Meenakshi
Nagarajan, Cartic Ramakrishnan,
Amit P. Sheth, I. Budak Arpinar
LSDIS Lab, Dept. of Computer Science.
University of Georgia
Athens, GA 30602-7404
(boanerg, bala, cartic, amit, budak)@cs.uga.edu

Li Ding, Pranam Kolari,
Anupam Joshi, Tim Finin
Department of Computer Science and Electrical
Engineering
University of Maryland, Baltimore County
Baltimore, MD 21250
(dingli1, kolari1, joshi, finin)@cs.umbc.edu

ABSTRACT

In this paper, we describe a Semantic Web application that detects Conflict of Interest relationships among potential reviewers and authors of scientific papers. This application discovers various ‘semantic associations’ between the reviewers and authors in a populated ontology to determine a degree of Conflict of Interest. This ontology is built by integrating entities and relationships from two social networks, namely “*knows*” from a FOAF (Friend-of-a-Friend) social network, and “*co-author*” from the underlying co-authorship network of the DBLP bibliography. We describe our experiences on development of this application in the context of a class of Semantic Web applications which have important research and engineering challenges in common. In addition, we present an evaluation of our approach for real-life COI detection.

Categories and Subject Descriptors

H.4.m [Information Systems Applications]: Miscellaneous;
H.3.4 [Information Storage and Retrieval]: Systems and Software - Information Networks; E.1 [Information Storage and Retrieval]: Systems and Software

General Terms

Algorithms, Experimentation

Keywords

Semantic Web, Social Networks, Conflict of Interest, Peer Review Process, Entity Disambiguation, Data Fusion, Semantic Associations, Ontologies, RDF

1. INTRODUCTION

Conflict of Interest (COI) is typically known as a situation that may bias a decision. It can be caused by a variety of factors such as family ties, business [31] or friendship ties, and access to confidential information. Detecting COI is necessary in many situations such as contract allocation, IPO (Initial Public Offerings) or company acquisitions, corporate law, peer review of scientific research papers or proposals, etc. Besides ensuring

impartial decisions, detection of COI is also critical where ethical and legal ramifications could be quite damaging to individuals or organizations. The underlying technical challenge is also related to the common *connecting-the-dots* applications that are found in broad variety of fields including regulatory compliance, intelligence and national security [18], and drug discovery [24].

In some cases it can be difficult to detect COI because of the lack of available information. However, in many other cases, there exists implicitly and/or explicitly information in the form of social networks such as those on the Web. For example, LinkedIn.com social network comprising a large number of people from information technology areas could be used to detect COI in situations such as IPO or company acquisitions. MySpace.com, Friendster, and Hi5 contain social network data that could substantiate COI in situations of friendship or personal ties. The list keeps growing, for example, Facebook.com (targeted towards college students) has recently begun expanding to include high-school students. Club Nexus is an online community serving over 2000 Stanford undergraduate and graduate students [1]. Yahoo! 360° and the acquisition of Dodgeball.com by Google are recent examples where the importance of social network applications is not only evident from the millions of users that some of them have but also in the (even hundreds of) millions of dollars they are worth.

Although social networks can provide data to detect COI, one important problem lies in the lack of integration among sites hosting them. Moreover, privacy concerns prevent such sites from openly sharing their data. Therefore, we chose publicly available social network data to address the challenge of COI detection. Our selection of public sources serves two purposes. First, they provide an opportunity to address the problem of integrating different social networks. Second, we can demonstrate real-world examples of the relevance of the problem of COI detection.

The data we used comes from bibliographic literature in Computer Science research. The DBLP bibliography (dblp.uni-trier.de/) provides collaboration network data by virtue of the explicit co-author relationships among people. We made the assumption that this collaboration network represents an underlying social network. As a second social network, we used a multitude of FOAF documents from the Web where the “*knows*” relationship is explicitly stated. The aggregation of such FOAF documents by means of the “*knows*” relationship results in a social network. Although we anticipated significant challenges while integrating the two networks, the effort needed in addressing this challenge surpassed our initial expectations. For

example, DBLP has different entries that in real-world refer to the same person, such as the case for “R. Guha” and “Ramanathan V. Guha”. Thus, the need for entity disambiguation (also called entity resolution, or reference reconciliation) will likely continue to be a fundamental challenge in developing Semantic Web applications involving heterogeneous, real-world data. We believe that this integration effort of two social networks provides an example of how semantic technologies, such as FOAF, contribute towards enhancing the Web.

The contributions of this paper are as follows:

- We bring together a semantic & semi-structured social network (FOAF) with a social network extracted from the collaborative network in DBLP. We explain the challenges involved with respect to large-scale entity disambiguation to achieve integration of different social networks (together with our results and findings for this task)
- We introduce semantic analytics techniques to address the problem of COI detection
- We describe our experiences in the context of a class of Semantic Web applications which have important challenges in common. We illustrate how an application that we developed for COI detection is a simple yet representative application of this class. The application is built around the scenario of peer review process. Thus, we demonstrate not only an application for COI detection but also shed some light on what it takes to develop this type of Semantic Web applications

2. MOTIVATION AND BACKGROUND

This paper intends to characterize common engineering and research challenges of building practical Semantic Web applications rather than contributing to the theoretical aspects of Semantic Web. In fact, many of us in academia have seen multifaceted efforts towards realizing the Semantic Web vision. We believe that the success of this vision will be measured by how research in this field (i.e., theoretical) can contribute towards increasing the deployment of Semantic Web applications [25]. In particular, we refer to Semantic Web applications that have been built to solve commercial world problems [26, 33, 34]. These include Semantic Search [16, 37], large scale annotation of Web pages [11] and commercialized semantic annotation technology [17], as well as applications for national security [35]. The engineering process it takes to develop such applications is similar to what we present in this paper. The development of a Semantic Web application typically involves a multi-step process as follows:

1. *Obtaining high quality data:* Such data is often not available. Additionally, there might be many sites from which data is to be obtained. Thus, metadata extraction from multiple sources is often needed [10, 23]
2. *Data preparation:* It typically follows after obtaining the data. Cleanup and evaluation of the quality of the data is part of data preparation
3. *Entity disambiguation:* This continues to be a key research aspect and often involves a demanding engineering effort. Identifying the right entity is essential for semantic annotation and data integration (i.e., [6])
4. *Metadata and ontology representation:* Depending on the application, it can be necessary to import or export data using standards such as RDF/RDFS and OWL. Addressing

differences in modeling, representation, encodings, etc. can require significant effort

5. *Querying and inference techniques:* These are needed as a foundation for more complex data processing and enabling semantic analytics and discovery (i.e., [4, 19, 21])
6. *Visualization:* The ranking and presentation of query or discovery results are very critical for success of Semantic Web applications. The users should be able to understand how inference or discovery is justified on the data
7. *Evaluation:* Often benchmarks or gold standards are not available to measure success of Semantic Web applications. Most frequently used method is comparing application output with results from human subjects

These challenges are discussed throughout this paper in the context of developing a semantic application that addresses the problem of COI detection. Figure 1 illustrates the multi-step process of building Semantic Web applications along with the steps involved in our approach for COI detection.

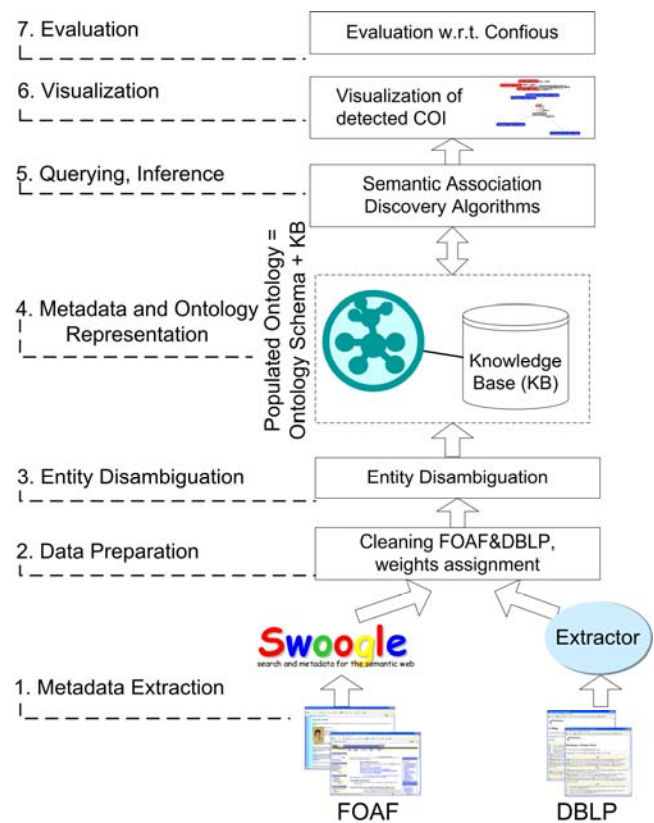


Figure 1. Multi-step Process of Semantic Web Applications

2.1 Conflict of Interest Detection Problem

Conflict of interest situations should be identified to produce impartial decisions, to comply with laws, etc. For example, the National Institute of Health (NIH), like many other government and private organizations, has strict definitions of what constitutes a COI. The NIH defines COI in the context of the grant review process as: “A Conflict Of Interest (COI) in scientific peer review exists when a reviewer has an interest in a grant or cooperative agreement application or an R&D contract proposal that is likely to bias his or her evaluation of it. A reviewer who has a real conflict of interest with an application or proposal may not

participate in its review.” Thus, one major cause for bias is professional or social relationships between potential reviewers and authors of the material to be reviewed. In this paper, we address the problem of COI detection in the context of peer review processes. We believe that the techniques presented here are applicable for COI detection in other scenarios.

2.2 The Peer Review Process

Throughout this paper, we will focus on peer review process for scientific research papers. This process is commonly supported by semi-automated tools such as conference management systems. In a typical conference, one (or more) person(s), designated as Program Committee (PC) Chair, is in charge of proper assignment of papers to be reviewed by PC members of the conference. Assigning papers to reviewers is probably one of the most challenging tasks for the Chair. State-of-the-art conference management systems support this task by relying on reviewers specifying their expertise, and/or “bidding” on papers. These systems can then assign papers to reviewers yet allowing the Chair to tweak these assignments. A key task is to ensure that there are qualified reviewers for a paper and that they will not have a-priori bias for or against the paper. These two requirements often conflict since publishing in top conferences is very competitive. Conference management systems can rely on the knowledge of the Chair about any particular strong social relationships that might point to possible COIs. However, due to the proliferation of interdisciplinary research, the Chair cannot be expected to keep up with the ever changing landscape of collaborative relationships among researchers, let alone personal relationships. Hence, conference management systems need to help the Chair for the COI detection task. Contemporary conference management systems support COI detection in different manners. EDAS (edas.info/doc/) checks for conflicts of interest based on declarations of possible conflicts by the PC members (i.e., while “bidding” for papers). Microsoft Research’s CMT Tool (msrcmt.research.microsoft.com/cmt/) allows authors to indicate COI with reviewers. Confius (www.confious.com) automatically detects these conflicts of interest based mainly on “similar emails” or “co-authorship” criteria. The “similar email” criterion tries to identify PC members and authors who are affiliated with the same organization based on the suffixes of the email addresses. The “co-authorship” criterion identifies users that have co-authored at least one paper in the past. However, Confius’ relatively straight forward approach can miss out on COIs as exemplified by one recent case. This particular case might have been undetected because the co-author in question now has a hyphenated last name. On the other hand, this is a good example of how difficult COI detection might be.

2.3 Online Social Networks

“A social network is a set of people (or organizations or other social entities) connected by a set of social relationships, such as friendship, co-working or information exchange.” [15]. Social networks are receiving a lot of attention on the Web because of an increasing number of websites allow users to post their personal information directly into online networked information spaces. The users of such websites form virtual or online communities. These have become part of the modern society in many contexts such as social, educational, political and business.

The entity *Person* is the fundamental concept in online social networks. An entity can be identified by one or several of its properties, and different sources might use different set of

properties, e.g. a person can be identified by his/her name in an office, but will be identified by his/her policy number by an insurance company. Such heterogeneous contexts and entity identifiers necessitate entity disambiguation. A “link” is another important concept in social networks. Some sources directly provide links among person entities such as `foaf:knows`, (where `foaf` refers to the FOAF namespace <http://xmlns.com/foaf/0.1/>). Other links, such as `co-author` among authors can be derived from metadata of publications.

Some of the online social networking sites provide machine readable personal information data using RDF/XML and FOAF vocabularies. Depending on the website’s privacy policy, the scope of published personal information ranges from nick names and interests to sensitive information like date of birth, etc. We acknowledge that there are privacy issues, yet a discussion on this is out of the scope of this paper.

2.3.1 Social Networks Analysis

Social network analysis focuses on the analysis of patterns of relationships among people, organizations, states and such social entities [7, 39, 40]. Social network analysis has applications in analysis of networks of criminals [41], visualization of co-citation relationships [8] and of papers [9], finding influential individuals [27, 36], study of the evolution of co-authorship networks [5], etc. Our work in this paper is fundamentally different than these previous approaches as it aims to develop and test a Semantic Web approach in integrating two social networks and using ‘semantic association’ discovery techniques for identification of COI relationships.

3. Social Network Integration for COI Detection

In order to demonstrate our approach to the problem of COI detection, we bring together a semi-structured yet semantic social network (FOAF) with a structured social network extracted from the underlying co-authorship network in DBLP. Here we describe these sources and explain the challenges involved with respect to entity disambiguation that have to be addressed to merge entities across (and within) these sources that in real-world refer to the same person.

3.1 Choosing Data Sources: FOAF and DBLP

We selected two representative online data sources for constructing two independent social networks, and then we combined them into one social network in the form of a populated ontology. These two real-world datasets were chosen based on the following criteria: first, they are representative for Semantic Web (FOAF) and database (DBLP) approaches; second, they consist of links among real-world persons, which is important for demonstrating COI detection; last, they are publicly available, thus facilitating their access with less privacy issues.

The Friend of a Friend (FOAF) data source, which is representative of Semantic Web data, is created independently by many authors because anyone can use the FOAF vocabulary to publish information about themselves and their social relationships. For example, a *person* entity can include identity-properties such as email and homepage, additional personal-properties such as name and personal photo using `foaf:name` and `foaf:depiction` respectively, and friendship-properties by means of `foaf:knows`. All this information can be encoded using an RDF/XML syntax thus making the corresponding social

network information machine readable. Many people maintain this type of social networks information in the FOAF world. For this reason, we can expect that people will use various sets of properties and that the values of such properties will be written using different conventions. The FOAF dataset we used [13], which includes 207,000 person entities based on 49,750 FOAF documents collected during the first three months of 2005, covers person entities in many professions, activities, etc. These FOAF documents are discovered by Swoogle Semantic Web search engine [12] (see item 1 in the multi-step process mentioned in Section 2).

The DBLP data source, which is representative of conventional database applications, is maintained by a single source. It is one of the best formatted and organized bibliography datasets. DBLP covers approximately 400,000 researchers who have publications in major Computer Science publication venues. Such bibliographic datasets have been used effectively to analyze collaborations [28], communities [38] and the spread of influence [22] in scientific communities. Most of this work has been aimed at understanding the structure of these networks. In DBLP, 'person' entities are fairly fixed – persons are identified by their names, and are associated by co-authorship relationships. Although counterexamples exist, such co-authorship relationships are well recognized as indicators of collaborative social relationships. Figure 2 illustrates the (ontology) schema and sample instances for the integrated network of DBLP and FOAF.

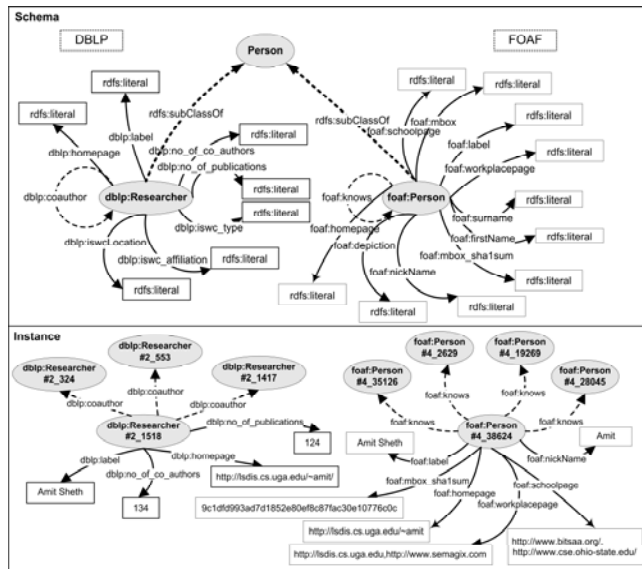


Figure 2. Schema and sample instances

3.2 Cleaning FOAF and DBLP Datasets

The goal of creating a combined dataset required maximizing the likelihood that DBLP entities will be connected to FOAF entities. Thus we chose authors of accepted papers in the 2004 and earlier International Semantic Web Conferences (ISWC) as well as the members of the Program Committee of these conferences as seeds. This set of people and their friends are likely to publish their personal profiles in FOAF and their names usually also appear in DBLP. We obtained two subsets from FOAF and DBLP as follows:

DBLP-SW: We collected 38,027 person entities that have up to three hops of social distance from those persons in Semantic Web (SW) conferences (as explained above). The following table

shows statistics of DBLP-SW (where 'dblp' is the alias for the namespace we used in this set):

Table 1. Statistics of DBLP-SW Dataset

Property	Persons having this relationship	
	#of entities	%
dblp:no_of_coauthor	38,015	99.96%
dblp:no_of_pub	38,015	99.96%
dblp:homepage	2,960	7.78%

FOAF-EDU: We first used the value of foaf:name to perform data cleaning. Examples of discarded names are "Tom's Website", "Shimone dot Org", as well as those containing special characters (i.e., '?', '{', '}'). This data-cleaning operation retained only about one third of the person entities (i.e. 66,112 instances). Then we applied several heuristics to identify researchers from person entities such as including FOAF documents residing on 'edu' websites. The following table shows statistics of FOAF-EDU, which contains 21,308 person entities:

Table 2. Statistics of FOAF-EDU Dataset

Property	Persons having this relationship	
	Quantity	%
foaf:mbox_sha1sum	14,169	66.50%
foaf:homepage	10,555	49.54%
foaf:nick	7,663	35.96%
foaf:depiction	5,016	23.54%
foaf:weblog	4,149	19.47%
foaf:firstName	2,913	13.67%
foaf:surname	2,865	13.45%
foaf:mbox	1,777	8.34%
foaf:workplaceHomepage	1,492	7.00%
foaf:schoolHomepage	766	3.59%

3.3 Entity Disambiguation

The types of Semantic Web applications exemplified by COI detection require high quality data. Hence it is necessary to resolve ambiguities among entities [32]. We adapted a recent work in name reconciliation [14] for resolving ambiguous entities in our datasets, and evaluated the effectiveness of this approach. We discuss our findings as we expect them to be applicable to this class of Semantic Web applications (see item 3 in the multi-step process mentioned in Section 2).

3.3.1 Disambiguation Algorithm

The goal is to find entities that might have multiple references in DBLP, FOAF, or both domains, which refer to the same entity (i.e., person) in real-life in order to establish a sameAs relationship between entities that are indeed the same entity (i.e., using owl:sameAs from W3C's OWL – Web Ontology Language). For this purpose, we adapted a name-reconciliation algorithm [14] that we selected for two reasons. First, it employs a rigorous form of semantic similarity by gleaning the context associated with an entity. Such similarity between two references is defined as a combination of the similarity between its atomic and association attributes (i.e., literal properties, and resource properties in RDF parlance). For instance, the reconciliation of two DBLP entities will be determined based on whether the similarity in their names, affiliations (atomic attributes) and the number of common co-author relationships (associations attributes) is sufficient evidence to reconcile the two entities with certainty (i.e., over a predefined threshold). The types of

relationships that an entity participates in thus have weights, which are assigned manually based on aspects such as their importance, the number of entities that have values for this attribute (see Tables 1 and 2), etc. The intuition behind weights (and thresholds) is to give more importance to relationships that define the context of an entity than to the syntactic similarity of attribute values. For example, relationships such as ‘homepage’ and ‘co-author’ contribute more contextual information than a ‘name’ relationship. Table 3 shows the weights of the attributes used in reconciling entities (in the two classes of entities of our dataset), as well as the merge thresholds. We found these weights and merge thresholds to be quite effective through several experiments where we considered multiple combinations of weights and merge thresholds values.

The second reason why we adapted the approach by Dong et al. [14] is because of the inherent nature of the data sources we used, where many entities lack enough information (i.e., attributes) to be utilized in disambiguation. This drawback is addressed by propagating reference-similarity information between reconciliation decisions and by enriching references of two reconciled entities. This provides additional information to be used in future reconciliation decisions (in subsequent iterations done as part of the algorithm). A description of further details of this algorithm is outside the scope of this paper.

Table 3. Atomic attributes weights and thresholds

Comparable Atomic Entities	Weights
Reconciling two FOAF entities	
<i>merge criteria: atomic attributes threshold > 0.5 and having at least 5 relationships to friends in common</i>	
Label	0.175
foaf:mbox_sha1sum	0.35
foaf:firstName	0.0875
foaf:surname	0.0875
foaf:homepage	0.05
foaf:webblog	0.05
foaf:mbox	0.05
foaf:nick	0.05
foaf:workplaceHomepage	0.05
foaf:schoolHomepage	0.05
Reconciling two DBLP entities	
<i>merge criteria: atomic attributes threshold > 0.6 and having at least 5 co-authors in common</i>	
Label	0.3
dblp:homepage	0.6
iswc:affiliation	0.1
Reconciling a FOAF and DBLP entity	
<i>merge criteria: atomic attributes merge threshold > 0.5 and at least 3 friends who are also in the co-authors list</i>	
foaf:label & dblp:label	0.2
foaf:firstName & dblp:label#firstName	0.15
foaf:surname & dblp:label#surname	0.15
dblp:homepage & foaf:homepage	0.25
dblp:homepage & foaf:workplaceHomepage	0.125
dblp:homepage & foaf:schoolHomepage	0.125

3.3.2 Entity Disambiguation Results

The output of our adapted algorithm populates two result sets – a “sameAs” set and an “ambiguous” set. The sameAs result set contains entity pairs identified as the same entity. The ambiguous set contains entity pairs having a good probability of being the same but without sufficient information to be

reconciled with certainty. Table 4 shows the properties of the dataset and the results obtained when we applied the reference reconciliation algorithm on the combined dataset. The entity pairs to be compared were selected based on syntactic similarity of their names.

Table 4. Properties of the dataset and results

Number of FOAF entities	38,015
Number of DBLP entities	21,307
Total number of entities	59,322
Number of entity pairs to be compared	42,433
Number of entity pairs for which a sameAs was established	633
Number of entity pairs compared yet without sufficient information to be reconciled	6,347

The lack of a gold standard prevented us from using precision and recall metrics (see item 7 in the multi-step process mentioned in Section 2). Instead, we measured statistics of false positives and false negatives by manually inspecting random samples of entity pairs from both the sameAs set and the ambiguous set. For each of these sets, we picked 6 random samples, each having 50 entity pairs. A false positive in the sameAs set indicates an incorrectly reconciled pair of entities and a false negative in the ambiguous set indicates a pair of entities that should have been reconciled, but were not. We found 1 false positive in the sameAs set, and 16 false negatives in the ambiguous set. We estimated with a confidence level of 95% that using this algorithm on this dataset, the number of false negatives in any ambiguous set will be between 2.8% and 7.8%. The number of false positives was estimated with the same level of confidence to be between 0.3% and 0.9%. We found the following as the most common reasons for false negatives:

- A pair of entities that should have been reconciled was not because of lack of enough number of attributes and only a partial match between the ones available. For example, two FOAF entities had only the label and mailbox-hash attribute available and they differed in the mailbox-hash value.
- Entity pairs under comparison have enough attributes for the algorithm to use, but have different values for multi-valued attributes like mailbox-hash or homepage.
- Entity pairs under comparison have a high similarity in atomic attribute values, but have very few association attribute matches. This was more prevalent in cases where the association attributes lists (i.e., co-authors and friends lists) were incomplete. The low similarity in association attribute matches cannot be discounted, because it is possible to have two DBLP entities that do not refer to the same real-world entity, but have a high similarity in comparable atomic attributes and a common co-author. For example, entities E1 and E2 in Figure 3 are DBLP instances like the one shown in Figure 2. Although they have a high similarity in attribute values and one co-author in common, they are really two different entities.



Figure 3. Entities with high similarity

- A pair of entities has very few attributes for comparison, but has a high match in the most semantically relevant attributes such as mailbox-hash or homepage. Because of the number of attributes available, their threshold was not high enough for them to be reconciled. For example, entities F1 and F2 in Figure 4 had only the label and homepage attributes available for comparison. Although the homepage attributes matched, there was not enough evidence to conclusively state reconciliation.

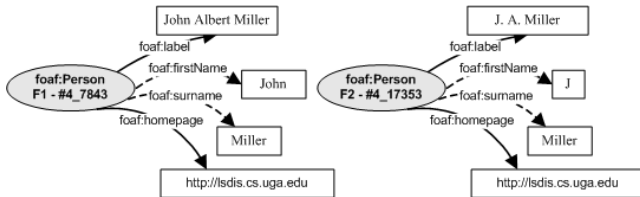


Figure 4. Entities with good match on relevant attributes

Although the objective of the implementation was to have as few false positives and false negatives as possible, we concluded based on experiments, that altering the weights and thresholds alone did not improve the results. The nature of the dataset, where a majority of entities appearing in FOAF have only between 3 and 7 attributes and the entities appearing in DBLP have between 3 and 5 attributes, plays a critical role in the results obtained. On the other hand, in cases like in Figure 4, showing FOAF entities F1 and F2, we found that it is possible to include data specific filters in the algorithm to obtain improvements on disambiguation results. For instance, a rule that specifies that two entity references should be considered the same if they have the same homepage or mailbox, in the absence of sufficient contextual information, might reduce the number of false negatives. However, such a consideration cannot be made without compromising on the results – if two people working in the same organization use the URL of their workplace as their homepage or misspell their email addresses, the use of such a filter would incorrectly identify them to be similar entities. Another way of improving results is whereby a conference management system using this method requests additional information such as affiliation, email, homepage, etc. (i.e., from authors of submitted papers) to be used in conjunction with the already available information.

4. SEMANTIC ANALYTICS FOR COI DETECTION

In this section we introduce different levels of COI and describe how we compute weights for relationships among the people in the integrated social network. Then we describe our algorithm for COI detection. This is followed by an experiment aimed at validating the ratings of various types of COI that our application identifies.

4.1 Levels of Conflict of Interest

By adhering to a strict definition of COI, there is only one situation in which there exists conflict of interest: the existence of a strong relationship. For other situations, an automated COI detection algorithm can provide insight by identifying potential COI. In this way, human involvement can be drastically reduced, but will still be relevant in cases where the quality of data is not perfect, the domain is not perfectly modeled, there is no complete data, etc. The subjective nature of the problem of COI detection is a good example where Semantic Web techniques cannot be

expected to be fully automatic in providing the correct solution. For these reasons, we introduce the notion of *potential* COI as it applies to cases where evidence exists to justify an estimated level of “low”, “medium”, or “high” degree of possible COI, as illustrated in Table 5.

Table 5. Levels of COI

Type	Level	Remarks
Definite COI	Highest	Sufficient evidence exists to require participant to abstain (i.e., recuse)
Potential COI	High	Evidence justifies additional verification of COI; participant is suggested to recuse
	Medium	Little evidence of potential COI
	Low	Shallow evidence of potential COI, in most cases it can be ignored

We now provide examples of each of the levels of COI mentioned above.

(1) “Definite” COI includes the case when a reviewer (i.e., PC member) is one of the listed authors in a paper to be reviewed (i.e., a reviewer must not review his/her own paper).

(2) “High” level of potential COI includes the existence of close or strong relationship(s) among an author of a submitted paper and a reviewer (i.e., a reviewer should not review the paper of a past collaborator).

(3) “Medium” level of potential COI occurs when a reviewer and an author of a paper to be reviewed have close relationships with a third party. For example, a reviewer may have a bias to an author if both had the same PhD advisor – even if they never collaborated or had any communication!

(4) “Low” level of potential COI includes situations with weak or distant relationships between a reviewer and an author of a paper to be reviewed. This degree of COI could, in most cases, be ignored. For example, “Anna” was PhD advisor of “John”, and “Beth” was PhD advisor of “Ken”, and the only relationship connecting “John” and “Ken” is through a co-authorship between “Anna” and “Beth”.

An algorithm for COI detection may also consider the quantity and strength of relationships to “upgrade” the level of COI (i.e., from “medium” to “high”). In addition, the ‘strenght’ of relationships should also be considered along with the ‘distance’ between a reviewer and an author to determine levels of COI. In our application, a preprocessing step computes weights for the strength of relationships among people in the integrated social network.

4.2 Weighting Relationships for COI Detection

A preprocessing step quantifies the relationships among people by their strength. This is particularly important for COI detection as the strength of relationships can facilitate validation of COI situations. The strength of relationships in the combined dataset is done by assigning weights between 0 and 1, where 1 refers to maximum strength. In our approach for COI detection, we assigned weights to two types of relationships, (FOAF) *knows* and (DBLP) *co-author*.

The relationship `foaf:knows` is used to explicitly list the friends a person has. These assertions can be weighted depending upon the provenance, quality, and/or reputation of their sources. On the other hand, the assertion of the `foaf:knows` relationship is

usually subjective and imperfect. For example, `foaf:knows` from A to B can be indicative of potential positive bias from A to B yet it does not imply a reciprocal relationship from B to A. Due to this, we assigned a weight of 0.5 to all 34,824 `foaf:knows` relationships in the FOAF-EDU dataset.

The second type of relationship we used for COI detection is the `co-author` relationship. Co-authorship is a good indicator for collaboration and/or social interactions among authors. However, we can find many counter examples against assumptions such as “one researcher always has a positive bias towards his/her collaborator” because friendship or positive opinion is not necessary for performing collaborative research. A more reasonable indicator of potential bias is the frequency of collaboration, which we use to assign weights to `co-author` relationships. For each researcher within DBLP-SW, we used the ratio of number of co-authored publications vs. total of his/her publications as the weight for the `co-author` relationship. That is, for any two co-authors a and b let $CO_{a \xrightarrow{\text{co-author-of}} b}$ represent the set of relationships where a co-authors a publication with b and the sets P_a and P_b represent the set of papers published by a and b , respectively. We define the weight of the `co-authorship` relationship between a and b as follows:

$$W_{a \xrightarrow{\text{co-author}} b} = \frac{|CO_{a \xrightarrow{\text{co-author-of}} b}|}{|P_a|}$$

Note that this weighting scheme is not symmetric. For example, the relationship `co-author` from “Li Ding” to “Tim Finin” has weight value of 0.5 because Ding has co-authored 3 papers with “Tim Finin” out of a total of 6 publications of Ding. On the other hand, the `co-author` relationship from Finin to Ding has a weight of 0.034 because Finin has co-authored many more publications (e.g., 87) with different collaborators. We collected 375,579 `co-author` relationships from DBLP-SW dataset. The weights computed for the relationships `foaf:knows` and `co-author` was represented using RDF reification and the combined dataset is serialized into RDF/XML to be suitable by our algorithms for discovery of semantic associations (see item 7 in the multi-step process mentioned in Section 2).

4.3 Detection of Conflict of Interest

Detecting the levels of COI listed in Table 5 requires analyzing the multiple ways in which two people are related. In our case, we had preliminary work on a particular type of queries, namely discovery of semantic associations [4]. This type of query is a capability directly applicable for COI detection as it involves discovery of path-sequences inter-connecting two entities. In addition, analysis of semantics [2] of such path-sequences is relevant for COI detection. We focused on finding rather short path-sequences. In other applications such as anti-money laundering, it is necessary to process longer paths [3]. Additionally, these path-sequences allow us to provide evidence of the explicit relationships involved in a level of COI. Obtaining these “path-sequences” by using currently available RDF query languages has disadvantages. For example, six queries are required to find all paths up to length two connecting two entities [20]. Making use of existing techniques, such as discovery of semantic associations, is an example of how our application fits with item 5 in the multi-step process mentioned in Section 2.

Our algorithm for COI detection works as follows. First it finds all path sequences between two entities. For the scenario of peer review process, one entity is the reviewer (i.e., PC member) and the other is an author of the paper to be reviewed. For the problem of COI detection, we find path-sequences containing up to 3 relationships. These are sufficient for the levels of COI listed in Table 5. Each path-sequence is analyzed by looking at the weights of the individual relationships involved within the path sequence. Since each path-sequence is analyzed independently of the others, each direction in relationships such as `co-author` is considered by the algorithm. We determined the thresholds by experimenting with several COI situations until we found appropriate correspondence with the levels of COI listed in Table 5. The following cases are considered:

- (i) Reviewer and author are directly related through an explicit relationship (either `foaf:knows` or `co-author`). Depending on customizable thresholds the assessments are: “high” for (at least one) relationship having weight on the range *medium-to-high* (i.e., weight ≥ 0.3); “medium” for (at least one) relationship having weight on the range *low-to-medium* (i.e., $0.1 \leq \text{weight} < 0.3$); and “low” for (at least one) relationship having a *low* weight (i.e., weight < 0.1).
- (ii) Reviewer and author are not directly related but they are directly related to (at least) one common person. Let us refer to this common person as an *intermediary*. Two cases will give an assessment of “medium”. In the first case, there are many (i.e., 10) such intermediaries in common. In the second case, the relationships connecting to the intermediary (i.e., one from the reviewer and another from the author) have weight on the range *medium-to-high* (i.e., weight ≥ 0.3). If neither of these two cases holds, then the assessment is “low”.
- (iii) Reviewer and author are not directly related but there are two *intermediaries*, one of each, that are related. In other words, the collaborators (or friends) of the reviewer and author have some tie. Then the assessment is “low”. It can be argued that this case is a very low level of potential COI, with which we agree. However, we wanted to keep this assessment in order to make sure that this case is also considered. A “low” level of COI can just be ignored for the scenario of peer review process, but in other COI scenarios, it might have some relevance.

In addition to the assessment of COI level, in some cases there exists a secondary assessment that is provided as well. For example, the assessment might have been “medium” but also a secondary assessment, “low”, might indicate a rare `co-authorship` relationship.

4.4 Application Prototype

Instead of providing a separate architecture diagram, we refer to Figure 1, which includes the core techniques or our application. We applied these techniques with the goal of bringing together different capabilities such as extraction and integration of social network data up to the point on which remains a semantic problem. We address the semantic problem by using techniques of discovery of ‘semantic associations’ as the basis for analysis of potential COI relationships. The integrated social network, represented as a populated ontology, allows us to exploit the relationships among entities, both for integration itself and for COI detection. A graph visualization of the different relationships (i.e., `co-author`, `foaf:knows`) carries such information of relationships thus allowing the user to inspect the cause of a given level of COI (see item 6 in the multi-step process of Section 2).

4.5 Experimental Results

In order to evaluate the effectiveness of our techniques we faced the problem of lack of similar results to compare with. Hence, we selected a subset of papers and reviewers from 2004 International World Wide Web Conference. This choice was motivated by the lack of any benchmark for evaluation of such results. We evaluated detection of COI, where human involvement is typically required to make final decisions.

The scenario on which we evaluated our approach included a subset of 15 PC members of the *Semantic Web track* and 10 of the accepted papers with topics related to the same track. The rationale for this selection is that researchers in this field would be more likely to have made available some of their information using FOAF. Table 6 lists the PC members and authors of (some of the) papers in our evaluation. The list shows only those co-authors for whom there was some level of COI detected. The algorithm does not detect COI in path sequences passing through other co-authors (thus eliminating redundant findings of COI). The different levels of COI detected are indicated on each cell, some of which have a primary, and a secondary level of COI. We compare our application with the COI detection of Confious conference management system [29]. Confious utilizes first and last name to identify at least one co-authored paper in the past (between reviewers and authors of submitted papers). Confious thus misses COI situations that our application does not miss

(because ambiguous entities in DBLP are reconciled in our approach). Confious detects previous collaborations and raises a flag of possible COI. Our approach provides additional information such as the level of potential COI as well as the reason. For example, our approach indicates that “Ian Horrocks” and “Alon Y. Halevy” have a “low” level of potential COI caused by a previous yet very rare co-authorship relationship among them. Finally, our approach is enhanced by the relationships coming from the FOAF social network, although in this particular evaluation, there was no situation where foaf:knows existed among two entities without co-author relationships.

We manually verified these COI assessments. While in most cases our approach validated very well, very few cases did not. We explain these 98324 cases. (i) false-negatives caused by lack of information: we found a FOAF document where “Daniel Schwabe” mentions “Stefan Decker” yet our dataset did not have this information; a targeted crawl could retrieve the latest FOAF documents. (ii) Few of the “low” assessments were caused by co-editing rather than co-authorship, such as in the case of “Ian Horrocks” and “Alon Y. Levy” appear as editors (among others) in the Proceedings of a Workshop in 1997. We believe that the assessment should still be “low” level of potential COI. However, data extraction could be improved in order to differentiate between co-authorship and co-editing relationships.

Table 6. COI Results

Authors:	Reviewers:	Karl Aberer	Sean Bechhofer	Mark Burstein	Isabel Cruz	Stefan Decker	Aldo Gangemi	R.V. Guha	Jeff Heflin	Ian Horrocks	Jane Hunter	M. Koubarakis	J. Mylopoulos	Wolfgang Nejdl	Guus Schreiber	Nigel R. Shadbolt
Dennis Quan			L3							<u>L2</u>						<u>L1</u>
Sean Bechhofer	<u>L1</u>	D	<u>L1</u>	<u>L1</u>	<u>L1</u>	<u>M1</u>				H						<u>L4</u>
Alon Y. Halevy		<u>L1</u>			<u>L2</u>	<u>L5</u>		<u>L1</u>		LR, <u>L6</u>		<u>L3</u>	<u>L3</u>	<u>L1</u>		
Wendy Hall	<u>L1</u>	MR,				<u>L1</u>				<u>L2</u>	<u>L1</u>		<u>L1</u>	<u>L1</u>	<u>L2</u>	MR, <u>L2</u>
Leslie Carr	<u>L1</u>	<u>L1</u>				<u>L1</u>				<u>M1</u>			<u>L1</u>	<u>L2</u>	<u>L2</u>	MR, <u>L7</u>
Timothy Miles-Board		M, <u>L1</u>				<u>L1</u>				<u>M1</u>						<u>L1</u>
Christopher Bailey		M, <u>L2</u>														<u>L2</u>
Daniel Schwabe					<u>L1</u>	LR, <u>L2</u>				<u>L1</u>			<u>L1</u>	<u>L2*</u>	<u>L1</u>	
m. c. Schraefel	<u>L1</u>	<u>M1</u>	<u>M1</u>				<u>L1</u>							<u>L1</u>		M, <u>L2</u>
Nigel Shadbolt	<u>L1</u>	<u>L4</u>				<u>L8</u>				<u>L1</u>					MR, <u>L1</u>	D
Nick Gibbins	<u>L1</u>	<u>M1</u>				<u>M2*</u>									<u>L2</u>	H
Hugh Glaser	<u>L1</u>	<u>L2</u>													<u>L1</u>	M, <u>L5</u>
Steve Harris		<u>M1</u>													<u>L1</u>	H

H: High level of potential COI caused by previous co-authorship/friendship between reviewer and author

M: Medium level of potential COI caused by a previous *low-to-medium* co-authorship/friendship between reviewer and author

MR: Medium level of potential COI caused by a previous yet *rare (i.e., occasional)* co-authorship/friendship between reviewer and author

MN: Medium level of potential COI caused by ‘N’ collaborators/friends in common between reviewer and author

LR: Low level of potential COI caused by previous yet *very rare* co-authorship/friendship between reviewer and author

LN: Low level of potential COI caused by ‘N’ collaborators/friends in common between reviewer and author

D: Definite COI because the reviewer is an author of the paper to be reviewed

*: Indicate that there were foaf:knows relationships in the path sequences connecting a reviewer and an author

Underlined: indicate those situations that would not be detected by Confious.

5. DISCUSSION

Our experience in the problem of COI detection leads us towards discussing the following three questions:

What does the semantic Web offer today (in terms of standards, techniques, and tools)? Technical recommendations such as RDF(s) and OWL provide the basis towards standard knowledge representation languages in Semantic Web. In addition, query languages (www.w3.org/TR/rdf-sparql-query/), path discovery techniques [4], subgraph discovery techniques [30], are examples of existing techniques for analytical access on RDF data. With respect to data, the FOAF vocabulary has gained popularity for describing content (i.e., 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web, www.w3.org/2001/sw/Europe/events/foaf-galway). On the other hand, semantic annotation has been proven scalable [11] and supported by commercial products [17] gaining wider use.

What does it take to build Semantic Web Applications today? As we have seen in addressing the problem of COI, building Semantic Web applications is not a trivial task. At the current stage, development of these applications can be quite time consuming. As much as the Semantic Web is promoting *automation*, there is a lot of effort required in terms of manual efforts and in customization of existing techniques. The goal of *full/complete automation* is some years away. Currently, quality and availability of data is often a key challenge given the limited number of high quality and useful data sources. Significant work is required in certain tasks such as entity disambiguation. Thus, it is not straight forward to develop Semantic Web Applications. We cannot expect to have all the components readily available to build Semantic Web Applications. Even if they are available, proving their effectiveness is a challenging job due to the lack of benchmarks. On the other hand, had the current advances not been available, some applications would not have been possible. For example, which other openly available social network other than FOAF could have been used? Then again, a number of tools are available today that can make the manual work less intensive. While conceptually there has been good progress, we are still in an early phase in the Semantic Web as far as realizing its value in a cost effective manner.

How are things likely to improve in the future? Standardization of vocabularies used to describe domain specific data is invaluable in building semantic applications. This can be seen in the bio-medical domain e.g. the National Library of Medicine's MeSH (Medical Subject Heading) vocabulary that is used to annotate scientific publications in the bio-medical domain. Further research in data extraction from unstructured sources will allow for the semi-automated creation of semi-structured data for specific domains (based on the vocabularies) on which analytic techniques can be applied to build semantic applications like the one described in this paper. Analytical techniques that draw upon graph mining, social network analysis and a vast body of research in querying semi-structured data, are all likely to facilitate the creation of such Semantic applications. We expect that benchmarks will appear. In future there should be a large variety of tools available to facilitate tasks such as entity disambiguation, annotation of documents, etc.

6. CONCLUSIONS AND FUTURE WORK

We presented how an application for Conflict of Interest Detection fits in a multi-step process of a class of Semantic Web

applications which have important research and engineering challenges in common. In the process of building this application we have identified some major stumbling blocks in building applications that leverage semantics. These can be grouped into: (i) Data related issues such as metadata extraction, metadata quality and data integration, and (ii) Algorithms and techniques that can leverage semantics. Thus, in the future we can expect increased attention in techniques and tools for metadata extraction, quality assessment and integration benchmarks. We described how our approach for COI detection is based on semantic technologies techniques and provided an evaluation of its applicability using an integrated social network from the FOAF social network and the DBLP co-authorship network. We provided details on how these networks were integrated. We believe that the value of Semantic Web applications can only be possible by leveraging the implicit and explicit semantics of data, such as social networks. A demo of the COI application is available (lstdis.cs.uga.edu/projects/semdis/coi/). Based on our experiences developing this application, we discussed what the Semantic Web offers today, what it takes to develop Semantic Web applications, and how are things likely to improve in the future.

7. ACKNOWLEDGMENTS

This work is funded by NSF-ITR-IDM Award#0325464 titled 'SemDIS: Discovering Complex Relationships in the Semantic Web' and NSF-ITR-IDM Award#0219649 titled 'Semantic Association Identification Knowledge Discovery for National Security Applications.' We also acknowledge use of Semagix Freedom (www.semagix.com) in support for this research.

8. REFERENCES

- [1] Adamic, L.A., Buyukkokten, O. and Adar, E. A Social Network Caught in the Web. *First Monday*, 8 (6).
- [2] Aleman-Meza, B., Halaschek-Wiener, C., Arpinar, I.B., Ramakrishnan, C. and Sheth, A.P. Ranking Complex Relationships on the Semantic Web. *IEEE Internet Computing*, 9 (3). 37-44.
- [3] Anderson, R. and Khattak, A., The Use of Information Retrieval Techniques for Intrusion Detection. *First International Workshop on Recent Advances in Intrusion Detection*, (1998).
- [4] Anyanwu, K. and Sheth, A.P., r-Queries: Enabling Querying for Semantic Associations on the Semantic Web. *12th International World Wide Web Conference*, (Budapest, Hungary, 2003), 690-699.
- [5] Barabási, A.-L. *Linked-The new science of networks*. Perseus Publishing, Cambridge, MA, 2002.
- [6] Bergamaschi, S., Castano, S. and Vincini, M. Semantic integration of semistructured and structured data sources. *SIGMOD Rec.*, 28 (1). 54-59.
- [7] Berkowitz, S.D. An introduction to structural analysis: The network approach to social research. Butterworth, Toronto, 1982.
- [8] Chen, C. Visualizing Semantic Spaces and Author Co-Citation Networks in Digital Libraries. *Information Processing Management*, 35 (3). 401-420.
- [9] Chen, C. and Carr, L., Trailblazing the literature of hypertext: author co-citation analysis (1989 - 1998). *Tenth*

- ACM Conference on Hypertext and hypermedia*, (Darmstadt, Germany, 1999), ACM Press, 51-60.
- [10] Crescenzi, V., Mecca, G. and Merialdo, P. RoadRunner: Towards Automatic Data Extraction from Large Web Sites *Proceedings of the 27th International VLDB Conference*, Morgan Kaufmann Publishers Inc., Rome, Italy, 2001.
- [11] Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R.V., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A. and Zien, J.Y., SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. *Twelfth International World Wide Web Conference*, (Budapest, Hungary, 2003), 178-186.
- [12] Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V. and Sachs, J. Swoogle: A Search and Metadata Engine for the Semantic Web *CIKM 2004*, Washington, DC, USA, 2004.
- [13] Ding, L., Finin, T., Zou, L. and Joshi, A. Social Networking on the Semantic Web. *The Learning Organization*, 5 (12).
- [14] Dong, X., Halevy, A. and Madhavan, J., Reference Reconciliation in Complex Information Spaces. in *ACM SIGMOD/PODS Conference*, (Baltimore, Maryland, 2005).
- [15] Garton, L., Haythornthwaite, C. and Wellman, B. Studying Online Social Networks. *Journal of Computer-Mediated Communication*, 3 (1).
- [16] Guha, R., McCool, R. and Miller, E., Semantic Search. in *Twelfth International World Wide Web Conference*, (Budapest, Hungary, 2003).
- [17] Hammond, B., Sheth, A. and Kochut, K. Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content. In Kashyap, V. and Shklar, L. eds. *Real World Semantic Web Applications*, Ios Press Inc, 2002, 29-49.
- [18] Hollywood, J., Snyder, D., McKay, K.N. and Boon, J.E. *Out of the Ordinary: Finding Hidden Threats by Analyzing Unusual Behavior*. RAND Corporation, 2004.
- [19] Horrocks, I. and Tessaris, S. Querying the Semantic Web: A Formal Approach. *First International Semantic Web Conference on The Semantic Web*, Springer-Verlag, 2002.
- [20] Janik, M. and Kochut, K., BRAHMS: A WorkBench RDF Store And High Performance Memory System for Semantic Association Discovery. *4th International Semantic Web Conference*, (Galway, Ireland, 2005), Springer.
- [21] Karvounarakis, G., Alexaki, S., Christophides, V., Plexousakis, D. and Scholl, M., RQL: A Declarative Query Language for RDF. *The 11th International World Wide Web Conference*, (Honolulu, Hawaii, USA, 2002), 592-603.
- [22] Kempe, D., Kleinberg, J.M. and Tardos, É., Maximizing the spread of influence through a social network. in *KDD 2003*, (2003), 137-146.
- [23] Laender, A.H.F., Ribeiro-Neto, B.A., da Silva, A.S. and Teixeira, J.S. A brief survey of web data extraction tools. *SIGMOD Record*, 31 (2). 84-93.
- [24] Laz, T., Fisher, K., Kostich, M. and Atkinson, M. Connecting the dots *Modern Drug Discovery*, 2004, 33-36.
- [25] Lee, Y.L. Apps Make Semantic Web a Reality *SD Times*, 2005.
- [26] Miller, E., The Semantic Web is Here. in *Keynote at the Semantic Technology Conference 2005*, (San Francisco, California, USA, 2005).
- [27] Nascimento, M.A., Sander, J. and Pound, J. Analysis of SIGMOD's CoAuthorship Graph. *SIGMOD Record*, 32 (3).
- [28] Newman, M.E.J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98 (2). 404-409.
- [29] Papagelis, M., Plexousakis, D. and Nikolaou, P.N. CONFIOUS : Managing the Electronic Submission and Reviewing Process of Scientific Conferences *6th International Conference on Web Information Systems Engineering*, New York, NY, USA, 2005.
- [30] Ramakrishnan, C., Milnor, W.H., Perry, M. and Sheth, A.P. Discovering Informative Connection Subgraphs in Multi-relational Graphs. *SIGKDD Explorations*, 7 (2). (to appear).
- [31] safeMinds.org. Something Is Rotten In Denmark, 2004.
- [32] Sayyadian, M., Shakery, A., Doan, A. and Zhai, C., Toward Entity Retrieval over Structured and Text Data. in *1st Workshop on the Integration of Information Retrieval and Databases (WIRD'04)*, (University of Sheffield, UK, 2004).
- [33] Sheth, A.P., Enterprise Applications of Semantic Web: The Sweet Spot of Risk and Compliance. in *IFIP International Conference on Industrial Applications of Semantic Web*, (Jyväskylä, Finland, 2005).
- [34] Sheth, A.P., From Semantic Search & Integration to Analytics. in *Dagstuhl Seminar: Semantic Interoperability and Integration*, (IBFI, Schloss Dagstuhl, Germany, 2005).
- [35] Sheth, A.P., Aleman-Meza, B., Arpinar, I.B., Halaschek, C., Ramakrishnan, C., Bertram, C., Warke, Y., Avant, D., Arpinar, F.S., Anyanwu, K. and Kochut, K. Semantic Association Identification and Knowledge Discovery for National Security Applications. *Journal of Database Management*, 16 (1). 33-53.
- [36] Smeaton, A.F., Keogh, G., Gurrin, C., McDonald, K. and Soding, T. Analysis of Papers from Twenty-Five years of SIGIR Conferences: What Have We Been Doing for the Last Quarter of a Century. *SIGIR Forum*, 36 (2).
- [37] Townley, J. The Streaming Search Engine That Reads Your Mind *Streaming Media World*, 2000.
- [38] Tyler, J.R., Wilkinson, D.M. and Huberman, B.A. Email as spectroscopy: Automated discovery of community structure within organizations. *cond-mat/0303264*.
- [39] Wasserman, S. and Faust, K. *Social network analysis: Methods and applications*. Cambridge University Press., Cambridge, 1994.
- [40] Wellman, B. Structural analysis: From method and metaphor to theory and substance. in Wellman, B. and Berkowitz, S.D. eds. *Social structures: A network approach*, Cambridge University Press, Cambridge, 1988, 19-61.
- [41] Xu, J. and Chen, H., Untangling Criminal Networks: A Case Study. *Intelligence and Security Informatics, First NSF/NII Symposium*, (2003), 232-248.