Scalable Semantic Analytics on Social Networks for Addressing the Problem of Conflict of Interest Detection

BOANERGES ALEMAN-MEZA
University of Georgia
and
MEENAKSHI NAGARAJAN
Wright State University
and
LI DING
Stanford University
and
AMIT SHETH
Wright State University
and
I. BUDAK ARPINAR
University of Georgia
and
ANUPAM JOSHI
University of Maryland, Baltimore County
and
TIM FININ
University of Maryland, Baltimore County

In this paper, we demonstrate the applicability of semantic techniques for detection of Conflict of Interest (COI). We explain the common challenges involved in building scalable Semantic Web applications, in particular those addressing connecting-the-dots problems. We describe in detail the challenges involved in two important aspects on building Semantic Web applications, namely, data acquisition and entity disambiguation (or reference reconciliation). We extend upon our previous work where we integrated the collaborative network of a subset of DBLP researchers with persons in a Friend-of-a-Friend social network (FOAF). Our method finds the connections between people, measures collaboration strength, and includes heuristics that use friendship/affiliation information to provide an estimate of potential COI in a peer-review scenario. Evaluations are presented by measuring what could have been the COI between accepted papers in various conference tracks and their respective program committee members. The experimental results demonstrate that scalability can be achieved by using a dataset of over 3 million entities (all bibliographic data from DBLP and a large collection of FOAF documents).

Categories and Subject Descriptors: H.4.m [**Information Systems Applications**]: *Miscellaneous*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software - *Information Networks*
General Terms: Algorithms, Experimentation
Additional Key Words and Phrases: Semantic Web, Social Networks, Conflict of Interest, Peer Review Process, Semantic Analytics, Entity Disambiguation, Data Fusion, Semantic Associations, Ontologies, RDF, DBLP, SwetoDblp

This research was supported by NSF-ITR Awards #IIS-0325464 and #0714441 titled ‘SemDIS: Discovering Complex Relationships in the Semantic Web.’

Authors' addresses: Boanerges Aleman-Meza, LSDIS Lab, Dept. of Computer Science, University of Georgia, Georgia 30602; Meenakshi Nagarajan, Kno.e.sis Center, College of Engineering and Computer Science, Wright State University, Ohio 45435; Li Ding, Knowledge Systems, AI Lab, Department of Computer Science, Stanford University, California 94305; Amit P. Sheth, Kno.e.sis Center, College of Engineering and Computer Science, Wright State University, Ohio 45435; I. Budak Arpinar, LSDIS Lab, Dept. of Computer Science, University of Georgia, Georgia 30602; Anupam Joshi, Department of Computer Science and Electrical

Engineering, University of Maryland, Baltimore County, Maryland 21250; Tim Finin, Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Maryland 21250.

1. INTRODUCTION

Conflict of Interest (COI) is a situation where bias can exist or be perceived based on the relationships or connections of the participants involved either explicitly or implicitly. The connections between participants could come from various origins such as family ties, business (e.g., safeMinds.org) or friendship ties and confidential information. Detecting COI is required to ensure "fair-play" in many decision-making situations such as contract allocation, IPO (Initial Public Offerings) or company acquisitions, corporate law and peer-review of scientific research papers or proposals. Detection of COI is also critical where ethical and legal ramifications could be quite damaging to individuals or organizations. The underlying technical challenges are also related to the common *connecting-the-dots* applications that are found in a broad variety of fields, including regulatory compliance, intelligence and national security [Hollywood et al. 2004] and drug discovery [Laz et al. 2004].

The detection of COI usually involves analysis of social network data, which is hard to obtain due to privacy concerns. We chose a special case of COI detection in the peer review process: detecting COI between computer science researchers. This case does not involve much of a privacy concern because researchers are open to expose their identity in published research (listing collaborators) and in their participation on the research community, e.g., as reviewers or organizers of conferences. Hence, social and collaborative information is widely published via various media such as magazines, journals and the Web. In particular, the advance of Web technologies has facilitated the access to social information not only homepages of persons and hyperlinks but also via many social networking sites.

Social networking websites attract more and more people to contribute and share information. For example, the LinkedIn social network comprises a large number of people from information technology areas and it could be used to detect COI in situations such as IPO or company acquisitions. MySpace, Friendster, Orkut and Hi5 contain data that could substantiate COI in situations of friendship or personal ties. The list keeps growing. Facebook was targeted to college students but it has expanded to include high-school students and now it is open to anyone. Club Nexus is an online community serving over 2000 Stanford undergraduate and graduate students [Adamic et al. 2003]. The creation of Yahoo! 360° and the acquisition of Dodgeball by Google are relatively recent examples where the importance of social network applications is evident not only considering the millions of users that some of them have but also due to the (even hundreds of) millions of dollars they are worth. Hence, it is not surprising that social network websites do not openly share their data. Other reasons for not doing so include privacy concerns. In some sites, the true identity of users is available only to their connections in the same network (e.g., Facebook, LinkedIn). Other sites such as LiveJournal publish the social connections of users openly yet the true identity of users is (in most cases) hidden behind a nickname.

Although social network websites can provide data to detect COI, they are isolated even when their users might overlap a lot. That is, many people have accounts in more than one site. Moreover, much of the social information is still hosted in the distributed homepage-hyperlink style. Therefore, our case of demonstrating COI detection faces a big challenge: integration of different social networks. Meanwhile, our case also serves as a real-world showcase of Semantic Web technology. The Friend-of-a-Friend (FOAF) vocabulary can be used to publish information about persons, their relationships to workplaces and projects, and their social relations. We used a collection 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. As a second network, we used the DBLP bibliography (dblp.uni-trier.de/), which provides collaboration network data by virtue of the explicit co-author relationships among authors. We made the assumption that this collaboration network represents an underlying social network. Although we anticipated significant challenges for the integration of the two networks, the effort needed in addressing this challenge surpassed our initial expectations. For example, DBLP has different entries that in the real world refer to the same person, such as the case of "Ed H. Chi" and "Ed Huai-hsin Chi." 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 to enhancing the Web.

This paper extends our previous work on semantic analytics on social networks [Aleman-Meza et al. 2006] where we demonstrated and explained the challenges of bringing together a semantic & semi-structured social network (FOAF) with a social network extracted from the collaborative network in DBLP. We also introduced semantic analytics techniques to address the problem of COI detection and described our experiences in the context of a class of Semantic Web applications where COI was a simple yet representative application. In this paper, our contributions go beyond those of the previous paper and can be summarized as follows.

- We verify scalability on bringing together a FOAF social network with the collaborative network in DBLP. We discuss the challenges in entity disambiguation to achieve integration of different social networks. Our evaluations demonstrate the need and feasibility of using large datasets (i.e., populated ontology with over 3 million entities).
- We improve upon our previous technique for COI detection by considering collaboration strength instead of basic co-authorship statistics. In addition, our new approach takes into account other relationships among people such as same-affiliation and co-editorship.
- We showcase the development process of creating scalable Semantic Web applications. Previously, we shed some light on what it takes to develop "connecting-the-dots" applications. Now we detail the challenges involved when, in addition, it is needed to use large-scale real-world datasets, in particular, social network data.

2. MOTIVATION AND BACKGROUND

This paper intends to characterize the common engineering and research challenges of building large-scale practical Semantic Web applications rather than contribute to the theoretical aspects of Semantic Web. In fact, many of us in academia have seen multi-faceted 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 to increasing the deployment of Semantic Web applications [Lee 2005]. In particular, we refer to Semantic Web applications that have been built to solve commercial world problems [Miller 2005; Sheth 2005a; Sheth 2005b]. These include Semantic Search [Guha et al. 2003; Wasserman and Faust 1994], large scale annotation of Web pages [Dill et al. 2003], commercialized semantic annotation technology [Hammond et al. 2002] and applications for national security [Sheth et al. 2005]. 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 the following multi-step process.

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 [Crescenzi et al. 2001; Laender et al. 2002; Sheth et al. 2002].

2. *Data preparation*: Preparation typically follows the obtaining of 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 (e.g., [Bergamaschi et al. 1999; Hassell et al. 2006]).
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 and encodings 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 (e.g., [Anyanwu and Sheth 2003; Horrocks and Tessaris 2002; Karvounarakis et al. 2002; Sheth et al. 2002]).
6. *Visualization*: The ranking and presentation of query or discovery results are very critical for the success of Semantic Web applications. Users should be able to understand how inference or discovery is justified by the data.
7. *Evaluation*: Often benchmarks or gold standards are not available to measure the success of Semantic Web applications. A 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 large-scale 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.

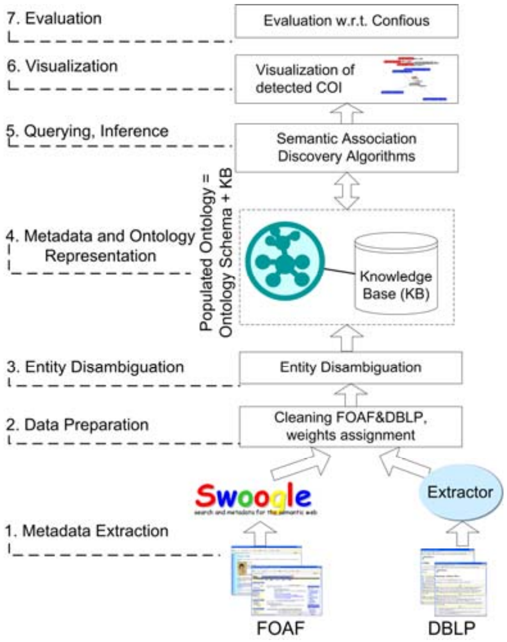
7. Evaluation
Evaluation w.r.t. Confious
6. Visualization
Visualization of detected COI
5. Querying, Inference
Semantic Association Discovery Algorithms
Populated Ontology = Ontology Schema + KB
4. Metadata and Ontology Representation
Knowledge Base (KB)
3. Entity Disambiguation
Entity Disambiguation
2. Data Preparation
Cleaning FOAF&DBLP, weights assignment
Swoogle
Extractor
1. Metadata Extraction
FOAF
DBLP

Fig. 1. Multi-step Process of Semantic Web Applications.

2.1 The Peer-Review Process

Throughout this paper, we will focus on the 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, (typically) one person designated as Program Committee (PC) Chair is in charge of the proper assignment of papers to be reviewed by PC members of the conference. Assigning papers to reviewers is 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 and also allow the Chair to modify these assignments. A key task is to ensure that there are qualified reviewers for a paper. In addition, it is necessary to ensure that the reviewers will not have a-priori bias for or against the paper. These two requirements often conflict due to the trade-off between the two aspirations. Namely, a qualified reviewer is expected to be completely unbiased yet s/he actually is a member of the same scientific community. 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 their personal relationships. Hence, conference management systems need to help the Chair with the detection of COIs.

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 (e.g., while bidding for papers). Microsoft Research's CMT Tool (cmt.research.microsoft.com/cmt/) allows authors to indicate COI with reviewers. Confious (confious.com) automatically detects 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, Confious' relatively straightforward approach can miss out on COIs as exemplified by one recent case of a co-author who now has a hyphened last name. On the other hand, this is a good example of how difficult COI detection might be. The approach presented in this paper makes use of social relationships to detect COI other than just based on co-authorship. This is possible by combining DBLP data and FOAF data through entity resolution, which is the main improvement over Confious.

2.2 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" [Garton et al. 1997]. Social networks are receiving a lot of attention on the Web due to the increasing number of websites that allow users to post their personal information directly into online networked information spaces. The users of such websites form virtual or online communities that 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. For example, 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`, 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 privacy policy of each website, the scope of published personal information ranges from nicknames and interests to sensitive information (e.g., date of birth). We acknowledge that there are privacy issues but a discussion on this topic is out of the scope of this paper.

2.1.1 Social Networks Analysis. Social network analysis focuses on the analysis of patterns of relationships among people, organizations, states, etc. [Berkowitz 1982; Wasserman and Faust 1994; Welman 1988]. Social network analysis has applications in analysis of networks of criminals [Xu and Chen 2003], visualization of co-citation relationships [Chen and Carr 1999] and of papers [Chen 1999], finding influential individuals [Nascimento et al. 2003; Smeaton et al. 2002], study of the evolution of co-authorship networks [Barabási 2002], etc. Our previous work in this respect demonstrated an ontological approach in integrating two social networks and using 'semantic association' discovery techniques for identification of COI relationships [Aleman-Meza et al. 2006].

3. INTEGRATION OF TWO SOCIAL NETWORKS

In order to demonstrate our approach to the problem of COI detection, we bring together a semantic social network (FOAF) with a 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 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 processable." 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. Moreover, quality issues can arise due to mistakes humans make on publishing their FOAF profile. For example, an earlier analysis of FOAF data [Ding et al. 2005b] lists cases where wrong data in "inverse functional" properties lead to incorrectly determining that two entities are the same.

We created our FOAF dataset [Ding et al. 2005a] using the 0.64 million FOAF documents collected by Swoogle [Ding et al. 2004] (see item 1 in the multi-step process of Section 2) during 2005-2006. These documents are published by 3,170 different websites such as livejournal.com (135,358 documents), bbc.co.uk (23,574 documents) and w3.org (112 documents) that cover various professions and activities. They contribute 85 million triples describing over 14 million person entities.

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 520K researchers who have publications in major Computer Science publication venues. Bibliographic datasets have been used for social network analysis, such as studying the structure [Newman 2001a] and the spread of influence [Kempe et al. 2003] in scientific communities. In DBLP, Person entities are fairly fixed; they are identified by their names and are associated by co-author relationships. Although counterexamples exist, co-authorship relationships are well recognized as indicators of collaborative social relationships.

In our previous work [Aleman-Meza et al. 2006] we used a subset of about 40K person entities from DBLP. This time we aimed at demonstrating large-scale applicability and hence we use all of DBLP data. There are available RDF conversions of DBLP data such as that by Andreas Harth (sw.deri.org/~aharth/2004/07/dblp/) and a D2R server publishing the DBLP bibliography in RDF (www4.wiwiss.fu-berlin.de/dblp/). However, we used our own SwetoDblp ontology of Computer Science publications [Aleman-Meza et al. 2007] where the main data source is DBLP data but also includes datasets of publishers and universities that enrich the data with relationships such as affiliation (based on the homepage DBLP authors, if available). SwetoDblp thus contains all metadata of publications in DBLP and it is being updated on a monthly basis and made available online (in RDF). Further details on SwetoDblp are outside the scope of this paper but are available [Aleman-Meza et al. 2007].

3.2 Data Preparation for FOAF and DBLP

The goal of creating a combined dataset led us to maximizing the likelihood that DBLP entities will be connected to FOAF entities. Thus, we selected a set of person names that appear in DBLP as authors of papers in major Web, Semantic Web, AI and Database Conferences, Journals and Workshops. This set of people and their friends are likely to publish their personal profiles in FOAF. We obtained a dataset of FOAF profiles by taking as starting point name matches of these person names (approximately 40K profiles).

Data-cleaning has been performed before using the FOAF dataset. Since 'mailto:' is the well-known prefix used by values of `foaf:mbox`, we corrected 1,035 mbox values that do not start with 'mailto:' and based on its value, we then added corresponding mbox_sha1sum triples. Table I shows usage statistics of some of the most popular properties in the cleaned FOAF dataset. We took a seed dataset of persons (that we call P0) and incrementally expanded it as follows. Based on the name and homepage information of the seeding person entities (P0), we found 482,336 relevant FOAF person entities (P1) each of which either having the same name or the same homepage with a seeding person. Then we found person entities (P2) known (through `foaf:knows` relationships) by an entity from either P0 or P1, and then found person entities (P3) who have same homepage, weblog, mbox, name, and mbox_sha1sum with P2 (these five properties are carefully selected – the first three are defined as inverse functional property by FOAF ontology, and the latter two are heuristic). The same propagation process was iterated until reaching P8.

Table I. Statistics of some properties in the cleaned FOAF Dataset

	Persons having this relationship	
Property	#of entities	percentage
rdfs:seeAlso	12,490,030	88.91%
foaf:knows	11,915,013	84.81%
foaf:interest	10,915,791	77.70%
foaf:nick	9,926,653	70.66%
foaf:weblog	9,180,701	65.35%
foaf:mbox_sha1sum	5,021,965	35.75%
foaf:name	4,176,365	29.73%
foaf:homepage	3,277,241	23.33%
foaf:mbox	928,842	6.61%

3.3 Entity Disambiguation

The class of Semantic Web applications exemplified by COI detection requires high-quality data. Hence, it is necessary to resolve ambiguities among entities. We adapted a recent work in name reconciliation 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 of Section 2).

Related Work. The goal of entity resolution is to identify two entities that refer to the same real world entity but are represented differently within the same or between two data stores. This problem (also called record linkage or de-duplication) is an important part of information integration and certainly not a new challenge. There has been plenty of work in resolving entities in two database schemas and later in XML. In each of these cases, the information exploited has included string and structural similarities between the schema elements, attributes and instance values. A comprehensive survey of efforts in this area and future directions is provided in [Winkler 1999]. More recent approaches take into account the edges of the underlying graph for evaluating similarities. Our approach to entity disambiguation is closest in spirit to such techniques [Neville et al. 2003; Kalashnikov et al. 2005; Jonyer et al. 2000] in terms of the underlying data model and information we exploit. Closest to our work is [Bhattacharya and Getoor 2006], where the authors use similar attribute and inter-entity relationships like 'co-author' relationships. In addition, they propose a clustering algorithm to perform the resolution. Our work differs in that we use attribute and relationship similarities and employ a combination of rules and thresholds to identify possible resolutions. The approach for entity resolution presented here is a direct adaptation of a previously published algorithm that we explain in more detail in the following section. Several approaches have also been suggested for entity resolution in the bibliographic domain which is also close to the domain considered in this work. Since, a comprehensive survey of each of them is outside the scope of this work, we point to the following resource that lists work in this area among other entity resolution approaches, www.ics.uci.edu/~rnuray/biblio.html.

3.3.1 Disambiguation Algorithm. The goal is to find entities (in DBLP and/or FOAF) that refer to the same entity (e.g, person) in real-life in order to establish an identity relationship between entities that are indeed the same entity (e.g., using `owl:sameAs` from W3C's OWL – Web Ontology Language). For this purpose, we adapted a name-reconciliation algorithm [Dong et al. 2005], which 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). In addition, the related entities are not already resolved, which necessitates the use of such an iterative algorithm. The second reason why we adapted the approach by Dong et al. [2005] is its applicability for the data sources we used, where many entities lack enough information (i.e., attributes) to be utilized for disambiguation. The algorithm addresses this drawback by propagating reference-similarity information between reconciliation decisions and enriching references of reconciled entities. Thus, additional information can be used in reconciliation decisions of subsequent iterations done as part of the algorithm. Although the description of the original algorithm is outside the scope of this paper, we discuss our architecture and some details surrounding the adapted algorithm.

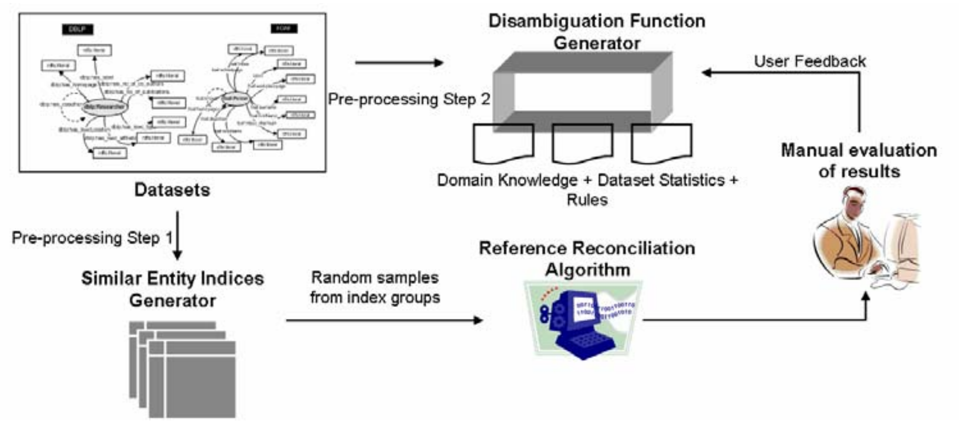
DBLP
FOAF
Datasets
Pre-processing Step 2
Disambiguation Function Generator
User Feedback
Domain Knowledge + Dataset Statistics + Rules
Manual evaluation of results
Pre-processing Step 1
Similar Entity Indices Generator
Random samples from index groups
Reference Reconciliation Algorithm

Fig. 2. Overall Design of the Entity Disambiguation System.

System Architecture. Figure 2 shows the overall design of our entity disambiguation system. Before the actual reconciliation of entities is attempted, we perform several pre-processing steps that contribute to the scalability of the algorithm. Our adapted reference reconciliation algorithm utilizes domain specific rules and a disambiguation function (Section 3.3.2 below) that takes as input a pair of entities that need to be reconciled and outputs a numerical score that indicates the confidence in the reconciliation. Additionally, indexes are created on the dataset to compare or reconcile only possibly similar entities. For example, there is no need to attempt to reconcile two entities with names ‘Amit P. Sheth’ and ‘Anupam Joshi.’ Section 3.3.3 discusses in more detail how we build our indexes. Once the disambiguation functions and rules are set up, sample entity pairs from the created indexes are run through the algorithm. A human evaluation of the reconciled entities alters the disambiguation function and rules until the user is satisfied with the results obtained.

The output of the adapted disambiguation 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. We now discuss each of the components of our architecture in more detail.

3.3.2 Disambiguation function and domain-specific rules. The core of the reconciliation algorithm is in designing a disambiguation function that can resolve whether two entities are considered to be the same in spite of data inconsistencies that may be present. The function uses attributes and relationships that an entity participates in and weights that indicate their contextual relevance to the reconciliation decision. Assigning weights to reflect the importance of attributes / relationships entails the use of statistics of the dataset (number of entities that have values for that attribute or relationship), some domain knowledge (co-authorship relationships of an author contribute more contextual information than an attribute value containing the number of his/her publications) and similarity functions (string similarity functions such as Levenshtein, Gotoh, and Jaro, Q-Grams). The weights shown in Table II were generated over 5 sample experiments (30 entity pairs in each) over the two datasets. Most of the samples were randomly chosen, while some used challenging disambiguation cases (often Asian authors with very similar names and insufficient attributes with which to compare) to be able to reach thresholds that are not too flexible or conservative.

Domain specific rules work on top of the disambiguation function deciding how much of the contextual information associated with an entity is to be utilized by the algorithm. Contextual information associated with an entity is the values of attributes and relationships the entity participates in. In many cases, just the atomic attributes like name, email addresses and affiliations suffice. In some other cases it is necessary to compare association attributes such as co-authorship relationships. Since utilizing association attributes (friends or co-authors) is more expensive, domain specific rules and thresholds are used to specify the order in which to consider attributes for disambiguation. Entity reconciliation proceeds by incrementally building contexts. In the process of reconciling entities, if there is enough information to reconcile two entities, the algorithm terminates the further accumulation of contextual information.

3.3.3 Building Indexes. When the dataset is large, comparing every entity with another is not reasonable. In order to optimize performance, we break our dataset into several buckets containing only entities that might be similar and need to be reconciled. Given our domain knowledge and dataset statistics, we create our indexes for the DBLP dataset based on the name and homepage attributes, and for the FOAF dataset based on the name and mailbox attribute values. In other words, if a group of entities have a high similarity in the aforementioned attribute values, they are placed in the same bucket to be reconciled. The thresholds used to decide if two entities need to be considered for reconciliation were reached via experiments and are very low (0.4 for DBLP entities, 0.5 for FOAF and 0.3 for DBLP and FOAF entities). Although a relaxed threshold results in a larger number of entities that need to be compared, it was important to not miss out on reconciling potentially similar entities at this stage.

3.3.4 Adapted Reference Reconciliation Algorithm. The reconciliation algorithm adapted from Dong et al. [2005] works by building and utilizing a context around every entity that needs to be disambiguated. Context specified by the important attribute and relationship values is built incrementally using domain specific rules (Section 3.3.3). One of the salient features of this algorithm is reusing past reconciliation decisions. Given the nature of our dataset where several entities have little or not enough attributes to use while performing the reconciliation, it was important to utilize as much information from past successful reconciliation decisions. When two entities are determined to refer to the same entity with a certain level of confidence, we merge all the information available to us thereby increasing the number of atomic and association attribute values available for the next reconciliation task. Albeit very effective, this step of the algorithm proved to be quite expensive when certain entities accumulated a large number of attributes over a series of reconciliation decisions.

Table II. Atomic Attributes Weights and Thresholds

Comparable Atomic Attributes	**Weights (old dataset)**	**Weights (new dataset)**
Reconciling two FOAF entities *Merge criteria: atomic attributes threshold > 0.5 and having at least 5 relationships to friends in common*		
Label (rdfs:label)	0.175	0.16
Foaf:mbox_sha1sum	0.35	0.17
foaf:firstName	0.0875	0.08
foaf:surname	0.0875	0.05
foaf:homepage	0.05	0.16
foaf:webblog	0.05	0.12
foaf:mbox	0.05	0.17
foaf:nick	0.05	0.14
foaf:workplaceHomepage	0.05	not used
foaf:schoolHomepage	0.05	not used
Reconciling two DBLP entities *Merge criteria: atomic attributes threshold > 0.6 and having at least 5 co-authors in common*		

Label (dblp:label)	0.3	0.3
Dblp:homepage	0.6	0.6
Dblp:affiliation	0.1	0.1
Reconciling a FOAF and DBLP entity *merge criteria: atomic attributes merge threshold > 0.5 and at least 3 friends who are also in the co-authors list*		
foaf:label & dblp:label	0.2	0.2
foaf:firstName & dblp:label#firstName	0.15	0.15
foaf:surname & dblp:label#surname	0.15	0.15
dblp:homepage & foaf:homepage	0.25	0.1667
dblp:homepage & foaf:workplaceHomepage	0.125	0.1667
dblp:homepage & foaf:schoolHomepage	0.125	0.1667

3.3.5 Entity Disambiguation Results. As mentioned previously, the output of the adapted disambiguation algorithm populates a sameAs result set containing entity pairs identified as the same entity, and an ambiguous set containing entity pairs having a good probability of being the same but without sufficient information to be reconciled with certainty. Table III 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. After data cleaning, some entities were removed which resulted in various components no longer connected to any of the names matching the initial list of person entities. Hence, we discarded a few large disconnected components of the FOAF dataset detailed in Table I and used approximately 580K person entities in the FOAF part of the integrated dataset.

Table III. Properties of the Dataset and Disambiguation Results

Number of FOAF person entities	579,030
Number of DBLP person entities	523,595
Total number of person entities	1,102,625
Number of entity pairs to be compared	40,592
Number of sameAs relationships established	2,387
Number of entity pairs that could not be reconciled (i.e., *ambiguous* relationships)	7,985

The lack of a gold standard prevented us from using precision and recall metrics (see item 7 in the multi-step process of 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. A false positive in *sameAs* is a pair of entities that were incorrectly identified as being the same. A false negative then is a pair of entities that were not identified as being the same, but actually were. 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 by 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%. In most cases, false negatives resulted because of the strict thresholds the algorithm applies. While one could vary these thresholds to obtain different results, these numbers worked well for the given dataset and application setting. It is hard to tell what is more expensive for such an application – a false positive or false negative in a *sameAs* result set. Since the goal is to make suggestions on conflicts of interest, we believe that it is more important to avoid false positives than false negatives in disambiguating two entities. Suggesting a COI because of incorrectly merging two entities, i.e., false positives, is different from an undetected COI which results from missing out on merging two entities, i.e., false negatives. Incorrect suggestions of both types require work on the user's part to identify and correct the COI. A false positive however also affects the algorithm in unfavorable ways. If two entities are incorrectly reconciled and their information (co-authors, publications etc.) is merged, it increases the risk of subsequently landing more incorrectly identified COIs. On the other hand, there is a possibility that a false negative fixes itself in subsequent iterations of the algorithm depending on the relationships connecting the co-authors. Given the many options that systems typically have for picking a reviewer, missing out on a few because of false negatives, is probably more tolerable than suggesting incorrect reviewers because of two entities reconciled incorrectly (i.e., false positive). The goal here was to therefore minimize the number of false positives. Since most of the false negatives were also due to insufficient information, it is very likely that they will be reduced if the algorithm had enough information.

We found the following as the most common reasons for false negatives:

- Entity pairs under comparison had a good number of attributes for the algorithm to use but with different values for their multi-value attributes. For example, two FOAF entities that were actually the same had the label, mailbox-hash and homepage attributes. The labels of such entities matched partially but the values of the mailbox-hash and homepage attributes differed. It is common for multi-value attributes like these to differ in values for two same entities. Problems arise when there are very few other non-multi-value attributes to use to establish confidence in the resolutions.

- Entity pairs under comparison had a high similarity in atomic attribute values, but had very few association attribute matches. This was more prevalent in cases where the association attributes lists (e.g., co-authors and friends) 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 that have a high similarity in attribute values and one co-author in common but are really two different entities.

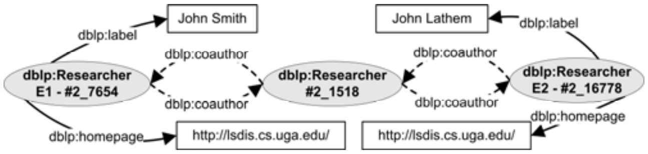

Fig. 3. Different Entities with High Similarity.

- A pair of entities that should have been reconciled was not due to insufficient attributes and having only a partial match between the attributes available. This is different from the first case, where there were sufficient attributes but were multi-valued in nature.

- A pair of entities had very few attributes for comparison, but had a high match in the most semantically relevant attributes such as mailbox-hash or homepage. Due to the small number of attributes available, their combined threshold was not high enough for them to be reconciled. For example, entities F1 and F2 in Figure 4 have very few attributes available for comparison. Although their homepage and surname attributes match, this is not enough evidence to conclusively state reconciliation.

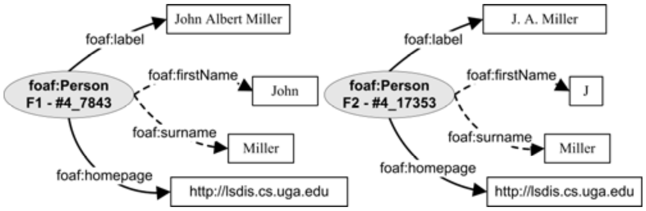

Fig. 4. Entities with Good Match on Relevant Attributes.

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 that of Figure 4, we found that it is possible to include data specific filters in the algorithm to obtain improvements on disambiguation results. For instance, a rule could specify that two entity references should be considered the same if they have the same homepage or mailbox in the absence of sufficient contextual information. However, such a consideration cannot be made without compromising on the results. For example, the case of two people using the URL of their workplace as their homepage would lead to incorrectly identifying them as the same entity. Another way of improving results is whereby a conference management system requests additional information such as affiliation, email and homepage (from authors of submitted papers) to be used in conjunction with the already available information (from FOAF and DBLP).

3.3.6 Scalability of the Adapted Algorithm. One of the key aspects of this experiment was also to evaluate how the algorithm scaled with respect to the new larger dataset. Below we present some of the challenges we faced and heuristics applied. The nature of the dataset also required us to change the threshold and weights of atomic and association attributes compared to the smaller dataset as shown in Table II.

- Use of past reconciliation decisions. This feature of the algorithm was very powerful for entities that lacked enough information. However, it occasionally caused overhead for entities that had enough contextual information that could be utilized. As heuristics, although we continued to merge information about entities, we started considering only the 5 most weighted atomic attributes before deciding if there was a need to use association attributes in the reconciliation. Further, when the confidence in the reconciliation surpassed thresholds shown in Table II, we consider the entities as the same and terminate the algorithm instead of trying to obtain the maximum confidence in reconciliation.

- Limit on the use of association attributes. In the previous version of our algorithm, in the event of weak or indecisive reconciliation decisions, we used multiple hops of association relationships to gather more context. For example, in disambiguating two entities from the FOAF dataset, we not only considered friends of the entities, but also friends of their friends as necessary. The number of hops used in the smaller dataset was 3 while we limited the number of hops to 1 for the larger dataset. Manual evaluation indicated that this did not affect the results since the nature of the dataset did not necessitate such wide spread information.

- Since the size of the individual datasets itself was large, we disambiguated references within the datasets (FOAF and DBLP) separately before finding same entities across the datasets. It is important to note however that since the resolution stage also enriches references i.e., combines all information known about the two resolved entities, subsequent steps of the algorithm in identifying references across the datasets remain unaffected.

4. SEMANTIC ANALYTICS FOR COI

In this section we review different levels of COI and describe how we computed weights for relationships among the people in the integrated social network. We then describe our updated algorithm for COI detection. This is followed by an experiment aimed at validating the various levels 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 a conflict of interest: the existence of a strong and direct 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 other cases, such as when the quality of data is not perfect, the domain is not perfectly modeled and when there is incomplete data. 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 use 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 IV. These are soft categories, which were created based on insight gained after examining the various ways in which two persons in the dataset can be (inter-) connected.

Table IV. Levels of Conflict of Interest (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/minor evidence of potential COI
	Low	Shallow evidence of potential COI, which in most cases can be ignored

We now provide examples of each of the levels of Table IV.

(1) “Definite” COI includes the case when a reviewer (e.g., 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 current/previous 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! This level of potential COI also occurs when a reviewer and an author have same affiliation. We have also included the case of editorship of a publication as Medium level of potential COI due to the different nature of co-authoring a paper. For example, many co-editor relationships come from putting together conference (or workshop) proceedings and hence do not involve the type of intellectual collaboration of co-authoring a publication.

(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, a reviewer and an author co-authored with a third party but it was not a significant collaboration. In fact, we no longer consider a three-degrees of separation as low conflict of interest as we did in our previous work.

4.2 Measuring Strength of Relationships

In our previous work, a preprocessing step quantified the strength of relationships between people. Weights were represented by means of reified statements. It has been noted elsewhere that the dataset size can drastically increase due to the verbosity of the XML serialization of RDF to represent reified statements [Mika 2005]. This would have an even larger impact on large datasets. An alternative was to use a quad store that deals with contextual information. In this work, we take a different approach that consists of computing strength of relationships at execution time. In addition, there are more types of relationships being considered for detection of COI. In our previous work, only two types were considered: FOAF *knows* and DBLP *co-author*.

The strength of relationships in the combined dataset is captured by weights between 0 and 1, where 1 refers to maximum strength. The relationship `foaf:knows` is used to explicitly list the person that are known to someone. 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 necessarily imply a reciprocal relationship from B to A. Hence, we assigned a weight of 0.45 to all `foaf:knows` relationships in the FOAF dataset. The cases where a `foaf:knows` relationship exists in both directions have a weight of 0.9.

Another type of relationship we used for COI detection is the *co-author* relationship, which is a good indicator for collaboration and/or social interactions among authors. However, counter examples can be found 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 compute weights of *co-author* relationships. In our previous work, we used the ratio of number of co-authored publications vs. total of his/her publications as the weight for the co-author relationship. However, such measure resulted in relatively low weights for co-authors of researchers that have many publications. For example, a researcher with over 100 publications had a very low co-authorship weight with a few of his doctoral students with whom had co-authored very few papers. Therefore, we now make use of a different measure of collaboration strength that takes into account the number of authors in a paper as well as the number of papers two people co-authored [Newman 2001b]. The formula adds a weight of 1/(n-1) to the collaboration strength of two authors for each paper they co-authored together (where n is the number of authors in a paper) [Newman 2001b]. This measure captures quite well the cases where a paper has very few authors based on the assumption that their collaboration strength is higher than in the case of papers with a large number of co-authors. The computed collaboration strength for any two co-authors is symmetric.

Lastly, another relationship among people that can be indicative of potential conflict of interest is that of 'same-affiliation.' The SwetoDblp dataset we used contains affiliation data for approximately 1% of people. Additional affiliation data comes from the FOAF part of the dataset but is not significantly larger than that of DBLP.

4.3 Detection of Conflict of Interest

Detection of levels of COI, as listed in Table IV, requires analysis of relationships between two persons. Hence, it is necessary to first discover and then analyze how two persons are connected by direct relationships or through sequences of relationships. Our previous work on discovery of "semantic associations" [Anyanwu and Sheth 2003] and their analysis [Aleman-Meza et al. 2005] is directly applicable for COI detection. This type of semantic analytics exploits the value of 'named' relationships and 'typed' entities with respect to an ontology. Thus, one of the benefits of an ontology-based approach for

COI detection is providing justification/explanation of the results by listing the semantic associations interconnecting the two persons. Semantic associations aim to capture meaningful and possibly complex relationships between entities in an ontology [Anyanwu and Sheth 2003]. Obtaining these semantic associations using currently available RDF query languages has disadvantages given that a semantic association is basically a path between two entities. For example, six queries are required to find all paths of up to length two connecting two entities [Janik and Kochut 2005]. In other applications, such as anti-money laundering, it is necessary to process longer paths [Anderson and Khattak 1998]. In our earlier prototype for COI detection, we looked for semantic associations containing up to 3 relationships. However, the dataset we use now requires us to find semantic associations containing up to 4 relationships. This is due to the fact that we did not preprocess the dataset to determine co-author relationships. Instead, the data contains implicit information about co-authorship in the form of two author entities being connected to a publication (by an intermediate RDF *blank node* that maintains the ordered sequence of the authors in papers). At execution time, semantic associations are reduced into shorter relationships such as co-author and same-affiliation (using some heuristics). The benefit of this is a level of abstraction whereby the COI detection algorithm is not concerned with representation details such as blank nodes. Hence, the work needed to adapt this application for usage of different datasets would not be significant. The utilization of existing techniques for complex data processing, such as discovery of semantic associations, is an example of how our application fits with item 5 in the multi-step process of Section 2.

Our algorithm for COI detection works as follows. First, it finds all semantic associations between two entities. For the scenario of peer-review process, one entity is the reviewer (e.g., PC member) and the other is an author of a paper to be reviewed. Second, each of the semantic associations found is analyzed to collapse it if applicable (as explained before) and then the strength of its individual relationships is computed. Since each semantic association is analyzed independently of the others, all directions of the different relationships are eventually considered. In our previous work, thresholds were required to decide between strong and weak collaborations. In our new implementation, a few combinations of relationships are analyzed:

(i) Reviewer and author are directly related (through `foaf:knows` and/or `co-author`). The assessments of potential level of COI is set to “high” regardless of the value of collaboration strength. The rationale behind this is that even a one-time collaboration could be sufficient reason for COI since it might have come from collaborating in a significant publication. Direct relationships through a same-affiliation relationship are given a "medium" potential COI level since it does not imply that the reviewer and author know each other. For example, some affiliation information is not up to date in the available data.

(ii) Reviewer and author are not directly related but they are related to one common person. Let us refer to this common person as an intermediary. Thus, the semantic association contains two relationships. An assessment of "medium" is set for the case where there are strong relationships connecting to the intermediary person. Otherwise, the assessment is set to "low." In the scenario of peer-review process, a low level of potential COI can be ignored but in other situations it might have some relevance.

We determined these cases and discarded our earlier method of using weight thresholds. The only exception is the case where a reviewer and author are not directly related but have strong relationship to an intermediary person. In addition to the assessment of COI level, our application keeps track of a secondary assessment (if any is detected), which is also shown to the user. For example, the assessment might have been "high" due to common co-authors but also could include a secondary assessment due to same-affiliation and/or co-editorship. The algorithm can be adapted to use different definitions than those of Table IV.

4.4 Evaluation Setup

Instead of providing a separate architecture diagram, we refer to Figure 1, which includes the core components of the application we built for evaluation. The goal was to bring together different capabilities, such as extraction and integration of social network data, up to the point on which it 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 representation of the data using an ontology, allows us to exploit the relationships among entities, both for integration and for COI detection.

As mentioned earlier, the dataset consisted of DBLP and FOAF data. The SwetoDblp ontology provided the DBLP data in RDF; we used the March-2007 version. It consists of metadata of over 800K publications, including over 520K authors thereof. The FOAF data consisted of about 580K persons linked through 519K `foaf:knows` relationships. The disambiguation process produced close to 2,400 relationships establishing *same-as* relationships in the integrated dataset. There are 4,478,329 triples between entities and

7,510,080 triples between entities and literal values. The dataset size in terms of disk space was of approximately 845 MB of DBLP data and 250 MB of FOAF data.

We utilized BRAHMS RDF database for building the prototype as it was designed for this type of connecting-the-dots applications [Janik and Kochut 2005]. BRAHMS creates a snapshot file for fast loading as main-memory database in subsequent usage. It took about 50 seconds to load our integrated dataset. All tests were performed on an Intel-based laptop with 2 GB of RAM running OSX 10.4. This shows that building this type of application is feasible without the need of expensive or sophisticated equipment such as dedicated servers or 64-bit architectures. The datasets used, the source code and the evaluation test cases (explained in the next section) are available online (lsdis.cs.uga.edu/projects/semdis/coi/).

4.4 Evaluation Setup

For the evaluation of the effectiveness of our techniques, we analyzed separately the accepted papers and Program Committee members of most tracks of the 2006 International World Wide Web (WWW) Conference. This choice was motivated by the lack of any benchmark for detection of COI, where human involvement is typically required to make final decisions. We selected this conference with the expectation that authors and reviewers in this field would be more likely to have made available some of their information using FOAF vocabulary. In addition, the organization of tracks in the WWW Conference facilitates evaluation due to their explicit grouping of papers per track where each track has a specific list of Program Committee members. The following are the different tracks in the conference that we used: Browsers & User Interfaces; Data Mining; E* Applications; Performance, Reliability and Scalability; Search; Security, Privacy and Ethics; and, Semantic Web. We do not include two tracks for which no COI cases were detected.

Tables V-XI list PC members and authors of the papers in our evaluation for which a potential COI was detected. We do not show the obvious cases of definite COI where a PC member is author of a paper. Also, we do not show cases of ‘low’ potential COI since in the scenario of peer-review these could be ignored. The tables show authors for whom there was some level of COI detected but does not list authors for which the COI depends on another author. For example, a doctoral student typically has published only with his/her advisor and any detected COI passes through connections of the advisor. The different levels of COI detected are indicated on each cell containing a primary, and in some cases, a secondary level of COI. We compared our application with the COI detection approach of the Confious conference management system [Papagelis et al. 2005]. Confious utilizes first and last names 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 For example, “Ed H. Chi” and “Ed Huai-hsin Chi” appear in different entries in DBLP yet they are the same person. Confious detects previous collaborations and raises a flag of possible COI. Our approach provides detailed information such as the level of potential COI as well as the cause. For example, our approach indicates that “Amit Sheth” and “Michael Uschold” have a “medium” level of potential COI due to co-editorship. In fact, each potential COI detected as “medium” or “low” would not have been detected by Confious. Finally, compared to Confious, the results of our approach are enhanced by the relationships coming from the FOAF social network. However, in the cases we tested there was no situation of two persons having a `foaf:knows` relationship and not having `co-author` or `co-editor` relationships between them. We expect that further adoption of FOAF will facilitate demonstration of its applicability. Nevertheless, conference management systems could ask authors and reviewers to provide their FOAF and make use of it directly for COI detection.

The key of cell values in tables is as follows:

D: Definite COI: reviewer is one of the authors

Hc: High potential COI: due to previous co-authorship

Mcc: Medium potential COI: due to common collaborator

Ma: Medium potential COI: due to same-affiliation

Me: Medium potential COI: due to previous co-editorship

Table V. COI Results - Browsers Track

WWW2006 Browsers Track	Krishna Bharat	Susan T. Dumais	Yoëlle S. Maarek	Paul P. Maglio	Andreas Paepcke	Dorée D. Seligmann	Terry Winograd
Prabhakar Raghavan	Hc	Hc	Hc		Ma		Ma
Alex Cozzi				Hc			
Jason Nieh						Ma	

Table VI. COI Results - Security Track (subset)

WWW2006 Security Track	Edward W. Felten	Kevin Fu	Craig Gentry	Benny Pinkas	Eric Rescorla	Brent Waters	Rebecca N. Wright
Dan Boneh	Hc	Hc	Hc	Hc	Hc	Hc,	Hc

					Mcc	Mcc, Ma	
John C. Mitchell						Ma	

Table VII. COI Results - Data Mining Track

WWW2006 Data Mining Track	Soumen Chakrabarti	Thomas Hofmann	Bing Liu	Wei-Ying Ma	Shinichi Morishita	Philip S. Yu
Chao Liu						Hc
ChengXiang Zhai		Hc				
Dou Shen				Hc		
Jian-Tao Sun				Hc		
Qiang Yang			Hc	Hc		
Zheng Chen				Hc		
Steven C. H. Hoi				Hc		
Michael R. Lyu				Hc		
Masaru Kitsuregawa					Ma	
Junghoo Cho	Hc					

Table VIII. COI Results - Performance Track

WWW2006 Performance Track	Martin F. Arlitt	Jeffrey S. Chase	Dejan Kostic	Michael Rabinovich	Oliver Spatscheck	Maarten van Steen	Alec Wolman
Balachander Krishnamurthy	Hc			Hc	Hc		
Craig E. Wills						Me	
Tracy Kimbrel		Hc					
Giovanni Pacifici				Hc			
Mike Spreitzer							Mcc
Patrick Reynolds		Hc	Hc				
Amin Vahdat		Hc	Hc, Mcc				

Table IX. COI Results - Search Track (subset)

WWW2006 Search Track	Junghoo Cho	Monika Henzinger	Panagiotis G. Ipeirotis	Anna R. Karlin	Christopher Olston	Sridhar Rajagopalan	Andrew Tomkins
Farzin Maghoul						Hc	Hc
Ravi Kumar						Hc, Mcc	Hc, Mcc
Ziv Bar-Yossef						Hc	Hc
Alexandros Ntoulas	Hc		Hc		Hc		
Marc Najork		Hc					
Mark Manasse				Hc			
Beverly Yang						Hc	
Soumen Chakrabarti	Hc				Hc	Hc	Hc

Table X. COI Results - E* Applications Track

WWW2006 E* Applications Track	John Domingue	Vincent P. Wade
Helen Ashman		Me
Amit P. Sheth	Hc	

Table XI. COI Results - Semantic Web Track (subset)

WWW2006 Semantic Web Track	V. Richard Benjamins	John Davies	John Domingue	Frank van Harmelen	Enrico Motta	Steffen Staab	Michael Uschold
Mustafa Jarrar						Hc	
Peter F. Patel-Schneider Ian Horrocks				Hc Hc		Hc	 Hc
Rudi Studer	Hc,Mcc	Me		Hc	Hc	Hc,Me	
Yolanda Gil	Me				Me		
Li Ding Amit P. Sheth Anupam Joshi Tim Finin	Hc		Hc			Hc Hc, Me Hc Hc	Me

We manually verified the COI assessments for the tracks listed. In most cases our approach validated very well but in rare cases it did not. For example, there is a 'high' level of potential COI between Amit Sheth and John Domingue due to co-authorship yet that particular case is from a 2-pages preface article in a Workshop organized by Drs. Sheth, Domingue and few others. A similar example is that of co-authors of Steffen Staab due to his IEEE Internet Computing column where one or more persons independently contribute with a section of the final article. In the resulting bibliography data of such articles, all authors appear as co-authors although they did not really collaborate as in regular research papers. These cases (Table XI) illustrate the dependency on the quality of the underlying datasets and/or data representation details. Similarly, we found few cases of false negatives that were due to various reasons: authors had co-authored a publication that was not listed in DBLP; and, authors had collaborated in a project but with no publication outcome yet. The lack of benchmarks for evaluation led us to select a few cases at random and visit the pages of authors on the web to look for false negatives. A comprehensive evaluation would require extensive manual effort. The fact that false positives were mostly due to lack of complete information, highlights the importance of obtaining high-quality data (Step 1 in Figure 1).

We noticed that some researchers have high potential COI with a number of other people. We looked into the data to glean a reason for this. We found that researchers having over 50 publications listed in DBLP data tend to show up more frequently in COI results. This is more noticeable for researchers with over 150 publications (examples in the tables listed include Drs. Ma, Raghavan, Sheth, and Staab).

In addition to the evaluation with respect to conference tracks and their respective papers, we created a list of persons that appear in FOAF to evaluate COI detection on the

FOAF part of the integrated network. We randomly selected 200 FOAF person entities that are connected to at least one other entity with a `foaf:knows` relationship. We evaluated them as fictitious authors and reviewers. Table XII illustrates a subset of the results that includes some researchers that also appear in the conference tracks listed before mentioned. The legend 'Mcf' indicates Medium potential level of COI due to common-friend; 'Lcf' indicates Low potential level of COI due to common-friend. The difference between Low and Medium rating for common-friend is that for Medium level it is necessary that the `foaf:knows` relationship exists in both directions (i.e., from A to B and from B to A).

Table XII. COI Results - Fictitious FOAF authors and reviewers

FOAF Person Entities appearing in WWW2006	Craig Knoblock	Tim Finin	Yimin Wang	Lalana Kagal	Jos de Bruijn	Emmanuel Pietriga	Marcelo Tallis
Pat Hayes	Mcf	Mcf		Lcf	Lcf	Lcf	Lcf
Cartic Ramakrishnan		Lcf	Mcf	Mcf			Lcf
Rong Pan			Mcf	Mcf			Lcf

We also measured the performance on COI detection excluding the time to load the dataset and dividing the remaining time by the number of pair-wise computations of COI detection (i.e., author and reviewer). On the average, it took 0.55 seconds to compute the COI between two persons. The majority of this is due to the search for the multiple semantic associations connecting them, of path length of up to 4 connections. Simple optimizations are possible such as starting the detection of COI with the authors that have published more papers. Figure 5 illustrates a sample screenshot that visualizes the resulting associations of two persons having collaborators in common. With respect to Step 6 in Figure 1, visualization provides a means for human to interpret the COI assessment.

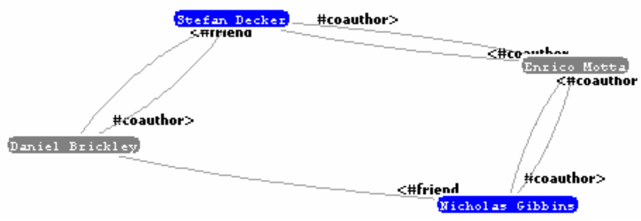

Fig. 5. Sample Screenshot of Visualization of Results.

5. DISCUSSION

In our previous paper [Aleman-Meza et al. 2006], we discussed three questions aiming at shedding some light on the efforts required and available tools/research to build semantic applications. Here we expand upon these issues by also addressing the scalability factor, which in many occasions is what determines whether research in the field is taking off.

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 [Anyanwu and Sheth 2003] and subgraph discovery techniques [Ramakrishnan et al. 2005] are examples of existing techniques for analytical access on RDF data, including recent developments that address extensions to SPARQL for expressing arbitrary path queries [Anyanwu et al. 2007; Kochut and Janik 2007]. With respect to data, the FOAF vocabulary has gained popularity for describing content (e.g., 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 [Dill et al. 2003] and supported by commercial products [Hammond et al. 2002] gaining wider use.

What does it take to build scalable Semantic Web Applications today? As we have seen by addressing the problem of COI, building scalable Semantic Web applications is not a trivial task. At the current stage, development of these applications can be quite time consuming. We demonstrated semantic techniques COI detection using real-world data. Existing open conference management systems (e.g., confious.org) could be extended to include the techniques presented here. 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. For example, a conference management system using techniques presented here should transparently choose the right dataset for COI detection regardless of whether the conference is a Computer Science conference or a Chemistry conference. 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 large-scale data sources. Significant work is required in certain tasks, such as entity disambiguation. Thus, it is not straightforward to develop scalable Semantic Web Applications because we cannot expect to have all the components readily available. Additionally, 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, which is used to annotate scientific publications in the bio-medical domain. Further research in data extraction from unstructured sources will allow semi-automated creation of semi-structured data for specific domains (based on the vocabularies) for 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 Semantic Web applications. We expect that benchmarks will appear. In the future, there should be a large variety of tools available to facilitate tasks, such as entity disambiguation and annotation of documents.

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, we identified some major stumbling blocks in building scalable applications that leverage semantics. These can be grouped into data related issues, such as metadata extraction, metadata quality and data integration as well as 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 demonstrated that scalability with respect to dataset size is possible. In addition, we highlighted the benefits of using a larger and richer dataset as well as using an improved method for COI detection. 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 application is available (lsdis.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 scalable Semantic Web applications and how are things likely to improve in the future.

In future work, we would like to handle the task of finding and assigning reviewers based on areas of expertise. Other systems use the social network of authors as basis to recommend reviewers (e.g., Referral Web [Kautz et al. 1997]). The contrast is that a referral system would be based on close/strong social connections yet these might bring along a conflict of interest.

ACKNOWLEDGMENTS

This work is funded by NSF-ITR Awards #IIS-0325464 and #0714441 titled 'SemDIS: Discovering Complex Relationships in the Semantic Web.'

REFERENCES

ADAMIC, L.A., BUYUKKOKTEN, O. AND ADAR, E. 2003. A Social Network Caught in the Web. *First Monday 8* (6).

ALEMAN-MEZA, B., HALASCHEK-WIENER, C., ARPINAR, I.B., RAMAKRISHNAN, C. AND SHETH, A.P. 2005. Ranking Complex Relationships on the Semantic Web. *IEEE Internet Computing 9* (3), 37-44.

ALEMAN-MEZA, B., NAGARAJAN, M., RAMAKRISHNAN, C., DING, L., KOLARI, P., SHETH, A.P., ARPINAR, I.B., JOSHI, A., FININ, T. 2006. Semantic Analytics on Social Networks: Experiences Addressing the Problem of Conflict of Interest Detection. In *Proceedings of the 13th International World Wide Web Conference*, Edinburgh, Scotland, May 2006, 407-416.

ALEMAN-MEZA, B., HAKIMPOUR, F., ARPINAR, I.B., SHETH, A.P. 2007. SwetoDblp Ontology of Computer Science Publications, *Journal of Web Semantics 5* (6). 151-155.

ANDERSON, R. AND KHATTAK, A. 1998. The Use of Information Retrieval Techniques for Intrusion Detection. In *Proceedings of the 1st International Workshop on Recent Advances in Intrusion Detection*, Louvain-la-Neuve, Berlin.

ANYANWU, K. AND SHETH, A.P. 2003. ρ-Queries: Enabling Querying for Semantic Associations on the Semantic Web. In *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary, 690-699.

ANYANWU, K., MADUKO, A. AND SHETH, A.P. 2007. SPARQ2L: Towards Support For Subgraph Extraction Queries in RDF Databases. In *Proceedings of the 14th International World Wide Web Conference*, Banff, Alberta, Canada.

ASWANI, N., BONTCHEVA, K. AND CUNNINGHAM, H. 2006. Mining Information for Instance Unification. In *Proceedings of the 5th International Semantic Web Conference*, Athens, GA, November 2006, 329-342.

BARABÁSI, A.-L. 2002. *Linked - The New Science of Networks*. Perseus Publishing, Cambridge, MA.

BHATTACHARYA, I. AND GETOOR, L. 2006. Entity Resolution in Graphs. In *Mining Graph Data*, L.B. HOLDER AND D.J. COOK, Eds. Wiley.

BERGAMASCHI, S., CASTANO, S. AND VINCINI, M. 1999. Semantic Integration of Semistructured and Structured Data Sources. *SIGMOD Record 28* (1), 54-59.

CHEN, C. 1999. Visualising Semantic Spaces and Author Co-Citation Networks in Digital Libraries. *Information Processing Management 35* (3), 401-420.

CHEN, C. AND CARR, L. 1999. Trailblazing the Literature of Hypertext: Author Co-citation Analysis (1989 - 1998). In *Proceedings of the 10th ACM Conference on Hypertext and Hypermedia: Returning to Our Diverse Roots*, Darmstadt, Germany, 51-60.

CRESCENZI, V., MECCA, G. AND MERIALDO, P. 2001. RoadRunner: Towards Automatic Data Extraction from Large Web Sites. In *Proceedings of the 27th International Conference on Very Large Data Bases*, Rome, Italy.

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. 2003. SemTag and Seeker: Bootstrapping The Semantic Web Via Automated Semantic Annotation. In *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary, 178-186.

DING, L., FININ, T., JOSHI, A., PAN, R., COST, R.S., PENG, Y., REDDIVARI, P., DOSHI, V. AND SACHS, J. 2004. Swoogle: A Search and Metadata Engine for the Semantic Web. In *Proceedings of the International Conference on Information and Knowledge Management*, Washington, DC.

DING, L., FININ, T., ZHOU, L. AND JOSHI, A. 2005a. Social Networking on the Semantic Web. *The Learning Organization 5* (12).

DING, L., ZHOU, L., FININ, T. AND JOSHI, A. 2005b. How the Semantic Web is Being Used: An Analysis of FOAF Documents. In *Proceedings of the 38th Hawaii International Conference on System Sciences*, Big Island, HI.

DONG, X., HALEVY, A. AND MADHAVAN, J. 2005. Reference Reconciliation in Complex Information Spaces. In *Proceedings of the ACM SIGMOD Conference*, Baltimore, Maryland.

GARTON, L., HAYTHORNTHWAITE, C. AND WELLMAN, B. 1997. Studying Online Social Networks. *Journal of Computer-Mediated Communication 3* (1).

GUHA, R., MCCOOL, R. AND MILLER, E. 2003. Semantic Search. In *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary.

HAMMOND, B., SHETH, A. AND KOCHUT, K. 2002. Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content. In *Real World Semantic Web Applications*, V. KASHYAP AND L. SHKLAR Eds. Ios Press Inc, 29-49.

HASSELL, J., ALEMAN-MEZA, B., ARPINAR, I.B. 2006. Ontology-Driven Automatic Entity Disambiguation in Unstructured Text. In *Proceedings of the 5th International Semantic Web Conference*, Athens, GA.

HOLLYWOOD, J., SNYDER, D., MCKAY, K.N. AND BOON, J.E. 2004. Out of the Ordinary: Finding Hidden Threats by Analyzing Unusual Behavior. *RAND Corporation*.

HORROCKS, I. AND TESSARIS, S. 2002. Querying the Semantic Web: A Formal Approach. In *Proceedings of the 1st International Semantic Web Conference*, Sardinia, Italy.

JANIK, M. AND KOCHUT, K. 2005. BRAHMS: A WorkBench RDF Store and High Performance Memory System for Semantic Association Discovery. In *Proceedings of the 4th International Semantic Web Conference*, Galway, Ireland.

JONYER, I., HOLDER, L. B., AND COOK, D. J. 2000. Graph-Based Hierarchical Conceptual Clustering. In *Proceedings of the 13th International Florida Artificial intelligence Research Society Conference*. AAAI Press, 91-95.

KALASHNIKOV, D., MEHROTRA, S. AND CHEN, Z. 2005. Exploiting Relationships for Domain-independent Data Cleaning. In *Proceedings of the SIAM Data Mining Conference*.

KARVOUNARAKIS, G., ALEXAKI, S., CHRISTOPHIDES, V., PLEXOUSAKIS, D. AND SCHOLL, M. 2002. RQL: A Declarative Query Language for RDF. In *Proceedings of the 11th International World Wide Web Conference*, Honolulu, HI, 592-603.

KAUTZ, H., SELMAN, B. AND SHAH, M. 1997. The Hidden Web. *AI Magazine 18* (2), 27-36.

KEMPE, D., KLEINBERG, J.M. AND TARDOS, E. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137-146.

KOCHUT, K., JANIK, M. 2007. SPARQLeR: Extended Sparql for Semantic Association Discovery, In *Proceedings of the 4th European Semantic Web Conference*, Innsbruck, Austria.

LAENDER, A.H.F., RIBEIRO-NETO, B.A., DA SILVA, A.S. AND TEIXEIRA, J.S. 2002. A Brief Survey of Web Data Extraction Tools. *SIGMOD Record 31* (2), 84-93.

LAZ, T., FISHER, K., KOSTICH, M. AND ATKINSON, M. 2004. Connecting the dots. *Modern Drug Discovery*, 33-36.

LEE, Y.L. 2005. Apps Make Semantic Web a Reality. *SD Times*.

MIKA, P. 2005. Flink: Semantic Web technology for the extraction and analysis of social networks. *Journal of Web Semantics 3* (2-3), 211-223.

MILLER, E. 2005. The Semantic Web is Here. (Keynote) In *Proceedings of the Semantic Technology Conference 2005*, San Francisco, California.

NASCIMENTO, M.A., SANDER, J. AND POUND, J. 2003. Analysis of SIGMOD's CoAuthorship Graph. *SIGMOD Record 32* (3).

NEVILLE, J., ADLER, M. AND JENSEN, D. 2003. Clustering relational data using attribute and link information. In *Proceedings of the Text Mining and Link Analysis Workshop*.

NEWMAN, M.E.J. 2001a. The Structure of Scientific Collaboration Networks. *Proceedings of the National Academy of Sciences 98* (2), 404-409.

NEWMAN, M.E.J. 2001b. Scientific Collaboration Networks: II. Shortest Paths, Weighted Networks, and Centrality, *Phys. Rev. E 64*, 016132.

PAPAGELIS, M., PLEXOUSAKIS, D. AND NIKOLAOU, P.N. 2005. CONFIOUS: Managing the Electronic Submission and Reviewing Process of Scientific Conferences. In *Proceedings of the 6th International Conference on Web Information Systems Engineering*, New York, NY.

RAMAKRISHNAN, C., MILNOR, W.H., PERRY, M. AND SHETH, A.P. 2005. Discovering Informative Connection Subgraphs in Multi-relational Graphs. *SIGKDD Explorations 7* (2), 56-63.

SHETH, A.P. 2005a. Enterprise Applications of Semantic Web: The Sweet Spot of Risk and Compliance. In *Proceedings of the IFIP International Conference on Industrial Applications of Semantic Web*, Jyväskylä, Finland.

SHETH, A.P. 2005b. From Semantic Search & Integration to Analytics. In *Proceedings of the Dagstuhl Seminar: Semantic Interoperability and Integration*, IBFI, Schloss Dagstuhl, Germany.

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. 2005.

Semantic Association Identification and Knowledge Discovery for National Security Applications. *Journal of Database Management 16* (1), 33-53.

SHETH, A.P., BERTRAM, C., AVANT, D., HAMMOND, B., KOCHUT, K. AND WARKE, Y. 2002. Managing semantic content for the Web. *IEEE Internet Computing 6* (4), 80-87.

SMEATON, A.F., KEOGH, G., GURRIN, C., MCDONALD, K. AND SODRING, T. 2002. 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).

TOWNLEY, J. 2000. The Streaming Search Engine That Reads Your Mind. *Streaming Media World*.

WASSERMAN, S. AND FAUST, K. 1994. *Social network analysis: Methods and applications*. Cambridge University Press., Cambridge.

WELLMAN, B. 1998. Structural analysis: From method and metaphor to theory and substance. In *Social Structures: A Network Approach*, B. WELLMAN AND S.D. BERKOWITZ Eds. Cambridge University Press, Cambridge, 19-61.

WINKLER, W.E. 1999. The State of Record Linkage and Current Research Problems. *RR99/03, US Census Bureau*.

XU, J. AND CHEN, H. 2003. Untangling Criminal Networks: A Case Study. In *Proceedings of Intelligence and Security Informatics, 1st NSF/NIJ Symposium*, 232-248.

Received ---; revised ---; accepted ---.