Analyzing Social Networks on the Semantic Web♦

Li Ding, Tim Finin, Anupam Joshi
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County

1 Introduction

The past year has seen a dramatic increase in the amount of social information published in RDF documents. Our investigations [1, 2] show that the Friend of a Friend (FOAF) ontology [3] is among the most used semantic web ontologies. This is true if we measure the number of semantic web documents (SWDs) that use the FOAF namespace, as Table I shows, or the number of triples using FOAF terms. The Swoogle Ontology Dictionary shows that the class *foaf:Person*[1] currently has nearly one million instances spread over about 45,000 web documents. The FOAF ontology is not the only one used to publish social information on the web. For example, Swoogle identifies more than 360 RDFS or OWL classes defined with the local name "person".

	namespace URI	amount of docs
1	http://www.w3.org/1999/02/22-rdf-syntax-ns#	200097 (96.9%)
2	http://purl.org/dc/elements/1.1/	146923 (71.2%)
3	http://purl.org/rss/1.0/	111595 (54.0%)
4	http://webns.net/mvcb/	68330 (33.1%)
5	http://xmlns.com/foaf/0.1/	49504 (24.0%)
6	http://www.w3.org/2000/01/rdf-schema#	44656 (21.6%)
7	http://purl.org/rss/1.0/modules/content/	28607 (13.9%)

Table I: The seven namespaces that were most frequently used in RDF documents known to Swoogle.

The Semantic Web and social network models support one another. On one hand, the Semantic Web enables online and explicitly represented social information; on the other hand, social networks, especially trust networks [4], provide a new paradigm for knowledge management in which users "outsource" knowledge and beliefs via their social networks [5]. In order to turn these objectives into reality, many challenging issues need to be addressed as the following.

- **Knowledge representation.** Although various ontologies capture the rich social concepts, there is no need to have hundreds of "dialectic" ontologies defining the same concept. How can we move toward having a small number of common and comprehensive ontologies?
- **Knowledge management.** The Semantic Web is, relative the entire Web, fairly connected at the RDF graph level but poorly connected at the RDF document level. The

♦ Partial research support was provided by DARPA contract F30602-00-0591(DAML) and NSF awards ITR-IIS-0326460 (SPIRE) and ITR-IIS-0325464 (SEMDIS). This is a draft of a short article to appear in IEEE Intelligent Systems (Trends & Controversies), volume 8, number 6, Nov/Dec 2004.
[1] it is the Qualified name (QName) of *http://xmlns.com/foaf/0.1/Person*.

open and distributed nature of the Semantic Web also introduces issues. How do we provide efficient and effective mechanisms for accessing knowledge, especially social networks, on the Semantic Web?

- **Social network extraction, integration and analysis.** Even with well-defined ontologies for social concepts, extracting social networks correctly from the noisy and incomplete knowledge on the (Semantic) Web is very difficult. What are the heuristics for integrating and fusing social information and the metrics for the credibility and utility of the results?
- **Provenance and trust aware distributed inference.** Provenance associates facts with social entities which are inter-connected in social network, and trust among social entities can be derived from social networks. How to manage and reduce the complexity of distributed inference by utilizing provenance of knowledge in the context of a given trust model?

2 Datasets

In order to understand how social networks on the Semantic Web are being modeled, we collected two datasets: DS-SWOOGLE and DS-FOAF[2]. The first dataset was collected by Swoogle [2] and provides a baseline model of the ontologies and information encoded in RDF on the Web. The dataset shows that the terms in the FOAF ontology, especially *foaf:Person*, are among the most used and populated[3] . We assume that it is reasonable to use the *foaf:knows* property to connect people forming social networks. Therefore, we collected the second dataset for the SemDis project [1] to focus on available FOAF documents containing instances of *foaf:Person*. Both datasets were collected from conventional web search engines, user-supplied URLs and our semantic web crawlers.

2.1 DS-SWOOGLE

At the time of this writing[4] , DS-SWOOGLE represents more than 225,000 valid Semantic Web Documents (i.e. online RDF documents in various formats such as 'RDF/XML' and 'N3') which contain about 37,000,000 RDF triples and are hosted by about 45,000 websites. Note that Swoogle samples at most 10,000 documents from each website to avoid being overwhelmed by websites with millions of RDF documents. Swoogle Ontology Dictionary and Swoogle Statistics are based on this dataset.

2.2 DS-FOAF and DS-FOAF-VAR

The DS-FOAF dataset collects URLs of over one million valid online FOAF documents[5] from over 1800 sites[6]. More than 95% of the URLs are hosted by five major 'blog' sites, which use limited vocabulary and fixed structure in describing personal profile. In order to reduce the impact from these sites, we studied a smaller datasets DS-FOAF-VAR that considers the websites which host at most 1000 FOAF documents. This dataset has over

[2]We noticed that these datasets are the largest ones among related works [1, 6, 7]
[3]A class or property is populated when it has instances. This is similar to the use of this word in databases.
[4]Swoogle is running continually and its database grows as new SWDs are added to the web
[5]We consider a FOAF document to be an RDF document that has at least one instance of the *foaf:Person* class
[6]We count web sites by DNS name and by IP address in DS-SWOOGLE and DS-FOAF respectively.

7,000 FOAF documents drawn from 1065 web sites that define nearly 37,000 instances of *foaf:Person*. These include 4,158 'strict' FOAF documents – ones intended to describe a single person and her acquaintances. Table II shows the detailed statistics of the two datasets.

	DS-FOAF			DS-FOAF-VAR		
	max	avg	std	max	avg	std
Persons /doc	2216	30.5	52.3	2196	5.1	49.4
SeeAlso /doc	2238	29.3	51.8	2066	1.9	36.7
Triples /person	-	-	-	3192	5.5	36.1

Table II: Statistics of DS-FOAF and DS-FOAF-VAR

3 Analytical Results

3.1 Building Common Social Ontology

One of the first practical contributions of the Semantic Web is that it provides a powerful distributed mechanism to represent and publish social network information. While the FOAF terms are widely used to encode social relations, other ontologies show up as well. We expect these to coalesce and merge as they evolve. In the light of the statistical approach to finding common terms [1, 6], we studied a particular class *foaf:Person*, which is the best used class in describing personal profile according to the statistics of DS-SWOOGLE and DS-FOAF. The definition of *foaf:Person* comes from three sources: (i) its *ontology definition* which relates it with other classes, (ii) the *ontological properties* that relate to it via *rdfs:domain* relation, and (iii) *empirical properties* that correlate with it by modifying its instances. An example is shown in Figure 1.

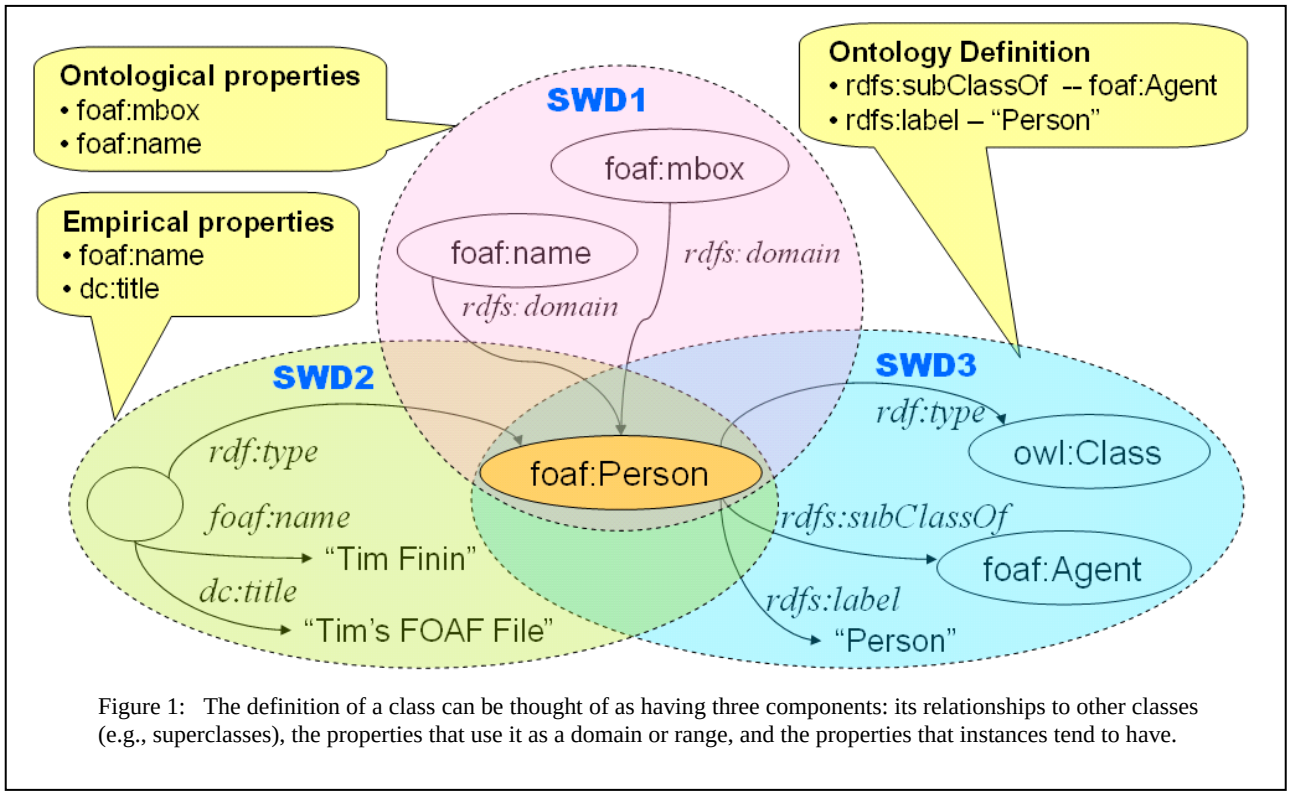

Figure 1: The definition of a class can be thought of as having three components: its relationships to other classes (e.g., superclasses), the properties that use it as a domain or range, and the properties that instances tend to have.

DS-SWOOGLE shows that *foaf:Person* has been defined by 17 ontologies. For examples, it is defined as both *owl:Class* and *rdfs:Class*; and has the named super-classes *foaf:Agent*, wordnet:Person, *http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing* and *http://www.w3.org/2000/10/swap/pim/contact#Person* DS-SWOOGLE reveals 162 ontological properties of *foaf:Person*, the majority of which represent social relations of one kind or another. There are also 74 properties whose *rdfs:domain* and *rdfs:range* are both *foaf:Person*. DS-SWOOGLE also finds 558 empirical properties of *foaf:Person* being populated with instance data. Table III and Table IV list 10 best used empirical properties, and suggest that people are concerned about privacy when publishing personal information: the property *foaf:mobx_sha1sum* is used much more frequently than *foaf:mbox*, hiding the true email address.

The empirical cardinality also shows that how users organize their profiles. The large value for max cardinality results from an unusual usage of FOAF vocabulary to build a collection of FOAF documents. In Table IV, the properties used frequently by documents but not by instances tend to be used to describe the owner of the strict FOAF documents.

Table III: Top 10 Empirical Properties of foaf:Person in DS-SWOOGLE

	property	max card	min card	docs	
				amount	percent
1	foaf:mbox_sha1sum	12	1	41403	95%
2	foaf:nick	7	1	36095	83%
3	foaf:weblog	5	1	35303	81%
4	rdfs:seeAlso	329	1	27838	64%
5	foaf:name	4	1	26749	62%
6	foaf:knows	3187	1	25736	59%
7	foaf:homepage	3	1	17616	41%
8	foaf:dateOfBirth	1	1	12783	29%
9	foaf:page	3	1	11255	26%
10	foaf:interest	300	1	10314	24%

Table IV: Top 10 Empirical Properties of foaf:Person in DS-FOAF-VAR

	prop usage per doc		prop usage per instance	
1	foaf:name	80%	foaf:name	65%
2	foaf:mbox_sha1sum	70%	foaf:mbox_sha1sum	60%
3	foaf:nick	51%	rdfs:seeAlso	37%
4	foaf:homepage	40%	foaf:nick	24%
5	foaf:depiction	35%	foaf:homepage	16%
6	foaf:weblog	30%	foaf:mbox	14%
7	foaf:knows	28%	foaf:weblog	14%
8	foaf:surname	27%	foaf:firstName	12%
9	foaf:firstName	27%	foaf:surname	12%
10	rdfs:seeAlso	26%	foaf:depiction	9%

3.2 Extracting Social Network

Extracting social network from noisy, real world data is a challenging task, even if the information is already encoded in RDF using well defined ontologies. The process consists of three steps: discovering instances of *foaf:Person*, merging information about unique individuals, and linking person through various social relation properties such as *foaf:knows*. A critical problem is determining whether two *foaf:Person* instances denote the same person. The semantics of FOAF vocabulary suggests several heuristics to answer this question:

- *named URI.* Non-anonymous individuals using the same URI denote the same person.
- *Inverse-functional properties.* Inverse functional properties such as *foaf:mbox* and *foaf:homepage* identify unique individuals. Other properties, such as *foaf:name* and *foaf:nick*, while not strictly inverse functional, can be used in practice in conjunction with other properties like *foaf:phone* to identify individuals with high probability.
- *Semantic equality.* When two or more values of an inverse functional property co-exist in the same individual's description, they are semantically equivalent as identifying the same individual.
- *rdfs:seeAlso.* This property almost always links to a strict FOAF document where the root person is the same as the referrer person.

In our preliminary study of DS-FOAF-VAR, we applied the first three heuristics and only consider *foaf:mbox_sha1sum* and *foaf:mbox* as inverse functional properties. We found 18,603 merged persons but only 10,247 of them have unique identifiers. Figure 2 shows cumulative distribution of the group size follows Zipf's distribution. Here, 'group' refers to the collection of individuals being merged as a person.

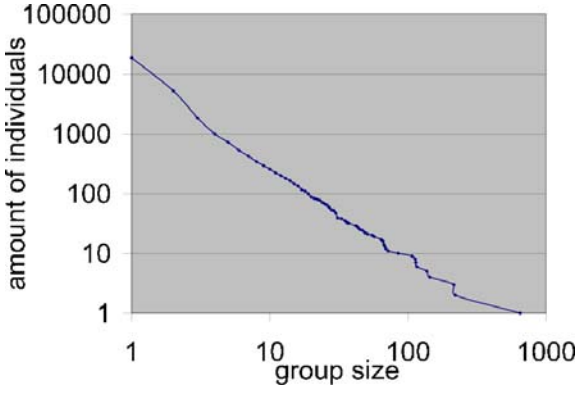

Figure 2: We merged individuals into groups based on several criteria. The distribution of the size of the equivalence groups follows a Zipf distribution.

These heuristics for merging individuals can fail in two distinct ways: inconsistency and separation. One *inconsistency* criterion is given by OWL, where cardinality constraints limit the semantically distinct values for a property. For example, when property P is restricted by having *owl:cardinality* one when modifying class C, all P's values in an individual of C should be semantically equivalent. In practice, according the common sense that "a person only has one name", we derive a cardinality constraint over *foaf:Person*. The semantic consistency of a person can be validated by checking whether it have two completely different names. *Separation* occurs when a person's information remains in two disjoint groups after merging. This gives rise to a dilemma – applying more merge heuristics may reduce separation but increase inconsistency.

3.3 Social network analysis

Social network analysis (SNA) is by itself a big research branch, our preliminary work limits in studying some basic graph features of the extracted social network. Mika [8] shows more applications of basic SNA measures on a smaller social network (n=167) extracted from FOAF and other web sources.

Degree analysis is one important measure in analyzing social networks. Our analysis of 14,164 distinct 'knows' relations in DS-FOAF-VAR shows that both in-degree and out-degree follow a Zipf distribution (Figure 3). We further put person into four categories: 'in only'(51.8%), 'out only'(5.8%), 'in+out'(5.4%), and 'isolated'(37.1%) according to their in-degree and out-degree. Such social network is not well connected because only a few ('in+out') persons are between the other persons plus that 94% 'in only' persons are known by only one person.

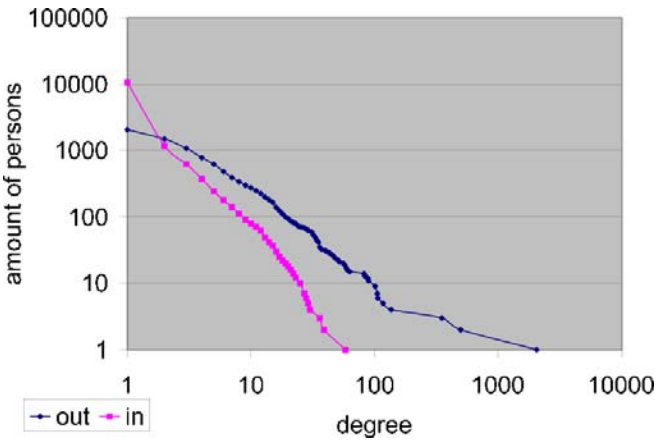

Figure 3: Cumulative distribution of in-degree and out-degree

Patterns of connected components. We have discovered 834 connected components and 6,904 isolated persons. The connected components exhibit interesting graphical patterns: (i) six *singletons* that link to themselves; (ii) a giant component which has 6,053 person connected; and (iii) several *stars* that have many out-links (the average out-degree for such nodes is 6.8). Figure 4 visualizes a selection of connected components. We hypothesize that the FOAF network topology evolves over time: a FOAF network starting from some disjointed star-alike connected components, then linking with one another to form trees and forests, eventually forming a scale free network.

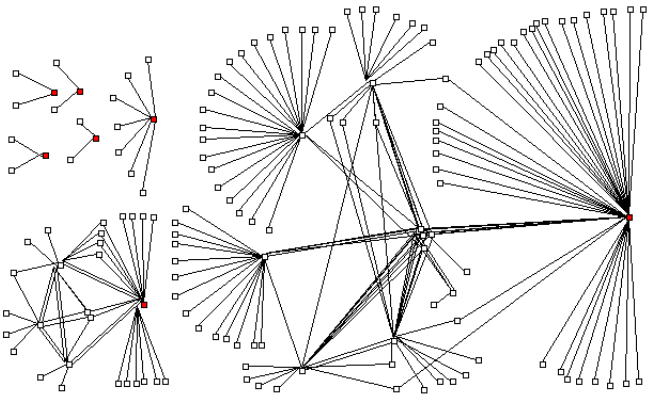

Figure 4: We found that connected components in FOAF networks followed a few simple patterns.

4 Conclusions

Our research is characterized by the following features: (i) it uses real world data in an open and distributed context, (ii) it provides data digest service for efficient data access on the Semantic Web, and (iii) it reasons over the knowledge encoded in semantic Web language. This paper describes research on integrating social ontologies and extracting social networks on the Semantic Web. We are also working on modeling trust across multiple social networks, and building a general architecture for provenance and trust aware distributed

inference in open, distributed and heterogeneous environments, such as the Web or multiagent systems.

Figure 5 illustrate our ongoing work on modeling trust across multiple social networks and reputations systems. In order to improve coverage and connectivity, we integrate social networks and reputation systems by mapping social entities like person. Then, trust relations maybe better derived and propagated through various social relations. As shown in Figure 5, the gap between “P. Kolari” and “A. Sheth” is connected by mapping “T. Finin” between two social networks. The reputation systems may offer default trust to social entities.

Our ongoing work is focused on continuing to improve the efficiency and effectiveness of data digest services, social network extraction and integration, and modeling provenance and trust for distributed inference services.

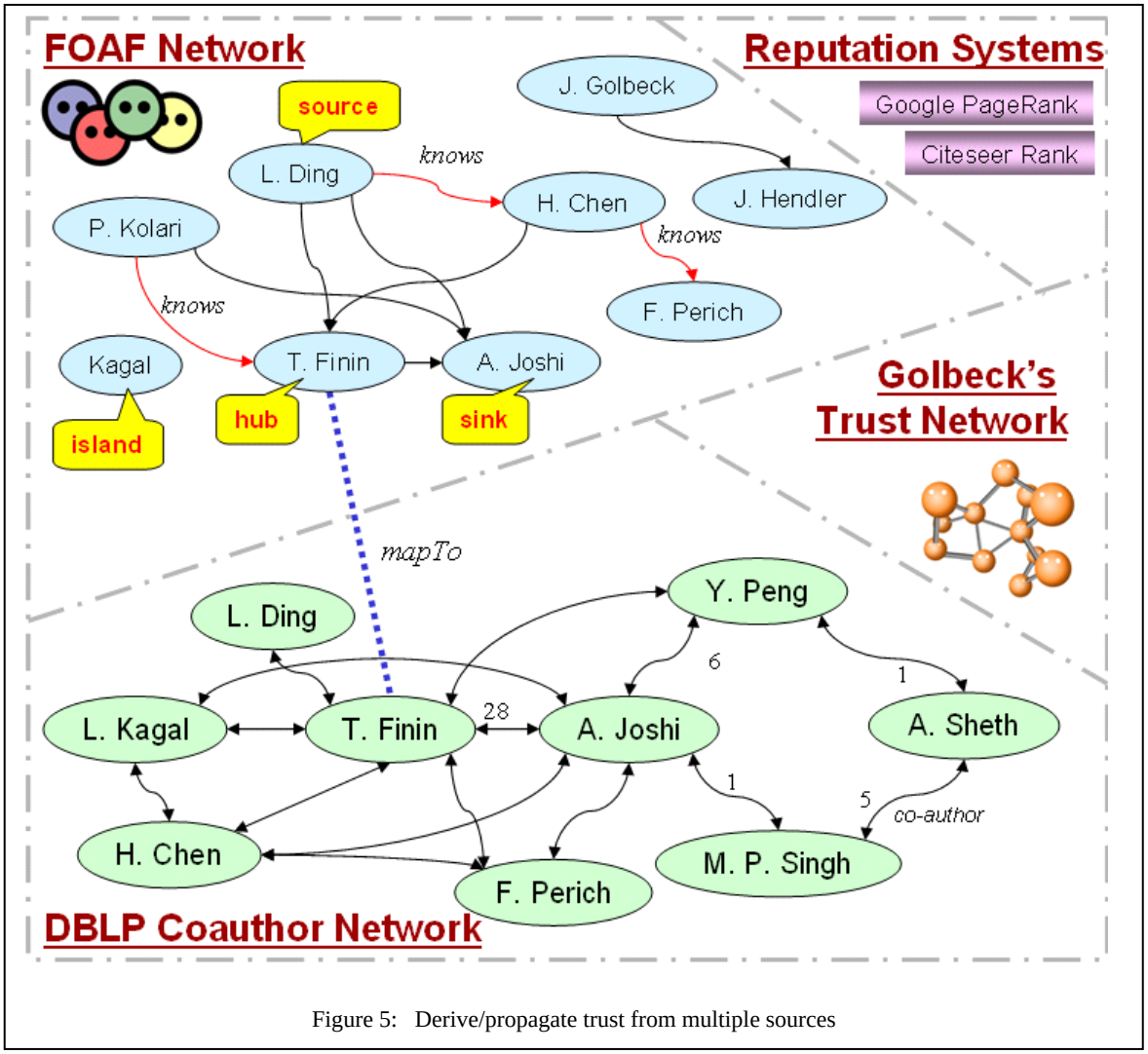

Figure 5: Derive/propagate trust from multiple sources

References

[1] L. Ding, L. Zhou, T. Finin, and A. Joshi, "How the Semantic Web is Being Used: An Analysis of FOAF," in Proceedings of the 38th International Conference on System Sciences, Digital Documents Track (The Semantic Web: The Goal of Web Intelligence), January 2005.

[2] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C. Doshi, , and J. Sachs, "Swoogle: A search and metadata engine for the semantic web," in Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management, Washington, DC, November 2004.

[3] "The friend of a friend (foaf) project," http://www.foaf-project.org/.

[4] J. Golbeck, B. Parsia, and J. Hendler, "Trust networks on the semantic web," in Proceedings of Cooperative Intelligent Agents, Helsinki, Finland , 2003.

[5] L. Ding, L. Zhou, and T. Finin, "Trust based knowledge outsourcing for semantic web agents," in Proceedings of IEEE/WIC International Conference on Web Intelligence, 2003.

[6] J. C. Paolillo and E. Wright, "The Challenges of FOAF Characterization," in Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web, 2004.

[7] G. A. Grimnes, P. Edwards, and A. Preece, "Learning meta-descriptions of the foaf network," in Proceedings of International Semantic Web Conference, Hiroshima, Japan, November 2004.

[8] P. Mika, "Bootstrapping the FOAF-Web: An Experiment in Social Network Mining," in Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web, Galway, Ireland, 1-2 September 2004,