Social networking on the semantic web

Tim Finin, Li Ding, Lina Zhou and Anupam Joshi
University of Maryland, Baltimore County, Baltimore, Maryland, USA

Abstract

Purpose – Aims to investigate the way that the semantic web is being used to represent and process social network information.

Design/methodology/approach – The Swoogle semantic web search engine was used to construct several large data sets of Resource Description Framework (RDF) documents with social network information that were encoded using the "Friend of a Friend" (FOAF) ontology. The datasets were analyzed to discover how FOAF is being used and investigate the kinds of social networks found on the web.

Findings – The FOAF ontology is the most widely used domain ontology on the semantic web. People are using it in an open and extensible manner by defining new classes and properties to use with FOAF.

Research limitations/implications – RDF data was only obtained from public RDF documents published on the web. Some RDF FOAF data may be unavailable because it is behind firewalls, on intranets or stored in private databases. The ways in which the semantic web languages RDF and OWL are being used (and abused) are dynamic and still evolving. A similar study done two years from now may show very different results.

Originality/value – This paper describes how social networks are being encoded and used on the world wide web in the form of RDF documents and the FOAF ontology. It provides data on large social networks as well as insights on how the semantic web is being used in 2005.

Keywords Internet, Information networks, Social networks

Paper type Research paper

1. Introduction

Social networks are explicit representations of the relationships between individuals and groups in a community. In the abstract, these networks are just simple graphs with nodes for the people and groups and links for the relationships. In practice, the links can encode all kinds of relationships – familial, friendship, professional or organizational. Social network theory, the study of such social networks, has developed techniques found useful in many fields, including sociology, anthropology, psychology and organizational studies. Social network analysis (SNA) has been used, for example, to represent and analyze the organization structure of employees in a business unit, identify key individuals, and suggest structural changes to improve unit performance.

Virtual or online communities are groups of people connected through the internet and other information technologies. These have become an important part of modern society and contribute to life in many contexts – social, educational, political and business. The communication technologies and infrastructures used to support virtual communities have evolved with the internet and include electronic mailing lists, bulletin boards, usenet, IRC, Wikis, and blogs. Virtual communities built on social

The Learning Organization
Vol. 12 No. 5, 2005
pp. 418-435

0969-6474
DOI 10.1108/09696470510611384

Partial support for this research was provided by DARPA contract F30602-00-0591 and by NSF awards NSF-ITR-IIS-0326460 and NSF-ITR-IDM-0219649.

network structures began appearing in 2002 and have become among most popular web-based applications. Such sites allow individuals to publish personal information in a semi-structured form and to define links to other members with whom they have relationships of various kinds. Current examples include Friendster, LinkedIn, Tribe.net, and Orkut. Other web-based virtual communities have successfully combined social networking with various interests, such as photography (www.Flickr.com), film (www.Netflix.com), personal blogging (www.Myspace.com) and dating (www.Thefacebook.com).

Several of these social network based virtual communities have begun to publish members' public profile information, including social links, using the semantic web language resource description framework (RDF). Most use the RDF vocabulary defined by the "Friend of a Friend" (FOAF) ontology augmented with new terms as needed. The use of a widely known, non-proprietary, shared ontology for this information enables interoperability among these systems. More importantly, RDF was designed as a data sharing standard privileging extensibility. Individual systems can extend the vocabulary as needed without interfering with the ability to combine and integrate information. This approach opens up many possibilities for information integration, aggregation and fusion on the web.

1.1 The semantic web and ontologies

The semantic web idea emerged from the confluence of several communities – artificial intelligence, hypertext, web developers – and so there are a number of ways to appreciate its motivation and goals. Perhaps the easiest for one who does not belong to any of those communities is to consider that much of what we want to know (that is actually known) is available on the web. Thus the web is, potentially, a great resource for software agents, which can be programmed to extract and fuse information from multiple, heterogeneous sources in response to a query.

However, extracting meaning from text is a very challenging task for computer programs. While progress is being made, a robust solution is decades, if not generations away. So the semantic web is an approach to encoding and publishing information in ways that makes it easier for computers to understand, thus making the web agent-friendly. What do we mean by "making it easier for computers to understand?" On the semantic web, we mean: through recourse to ontologies, formal descriptions of particular domains.

Ontology is the branch of philosophy that seeks to answer the question "what is there?". In computer science, an ontology is a formal conceptualization of a domain. Typically, it specifies the classes of objects that exist, the relationships amongst those classes, the possible relationships amongst instances of the classes, and constraints over those instances. An ontology also defines terms denoting these classes and relationships as well as individual objects. Current web ontology languages, designed to encode information on and for the web, use the eXtensible Markup Language (XML) both for specifying ontologies, and also for making assertions about the world using terms defined in ontologies. A semantic web page begins by listing (as URLs) the locations of the ontologies to be used, then goes on to use those ontologies to make assertions about datasets, human beings, items for sale, etc. An agent, on coming to such a page, can import the specified ontologies and use that information to understand the semantics of the ensuing assertions.

The world wide web consortium (W3C) has developed standards to enable ontologies to be published on the web as well as data and other assertions to be encoded using terms drawn from any published ontologies. These standards make it possible for programs and software agents to understand information published on the web without the ambiguity and complex processing inherent in traditional unstructured forms (e.g. natural language) or rigidity and lack of flexibility inherent in structured representations (e.g. relational databases).

The RDF (Klyne and Carroll, 2004) is a simple XML-based language to define computer-understandable vocabularies that people and programs can use to describe things of interest, such as web sites, newspaper articles, e-mail messages, people, books, events, or web services. RDF mimics human languages in that it allows one to introduce new terms (individuals, classes and properties) that are defined (partially, at least) in terms of existing terms. RDF Schema (Hayes, 2004) extends RDF by providing vocabulary to build logical object-oriented schema, including a simple typing system, sub-classes, sup-properties, inheritance, etc. The Web Ontology Language (OWL) (Schreiber and Dean, 2004) supports advanced capabilities, such as logical inference and translating descriptions using different ontologies (e.g. mapping a location specified as a ZIP code to one using latitude and longitude.)

A problem in the effort to formalize (or "ontologize") a domain is that there are typically many different ways of doing so. This is true whether the domain is in a science, or business-related or has to do with people and their relationships. Within a single discipline, there can be disagreement about how to describe the world. As well, disciplines overlap, and often look at the overlapping area from different points of view. One approach to the ontology heterogeneity problem is to create a global schema to serve as an interlingua for human and software agents. One of the principles of the semantic web is that it should be based on the same open, decentralized and distributed approach that has made the world wide web successful. Anyone should be able to create, publish and use their own ontologies. Mechanisms are available to allow one to define mappings or translations of terms among ontologies. In the open and dynamic environment of the web, it is expected that the natural influences and forces of the market and "networking effect" will encourage coalescing to a smaller number of interoperable ontologies for a given domain. So the construction of a few global schemata is not the goal. Rather, we envision and are encouraging the development of a number of relatively small ontologies, some of which may overlap, and some of which may be in conflict.

1.2 The FOAF ontology

The FOAF vocabulary includes classes and properties found useful to describe people online. Consider the following example, drawn from the FOAF vocabulary specification (www.foaf-project.org/) and encoded using the XML serialization for RDF.

```
< foaf:Person >
  < foaf:name > Dan Brickley < /foaf:name >
  < foaf:mbox_sha1sum > 241021fb0e6289f92815fc210f9e9137262c252e < /foaf:mbox_sha1sum >
  < foaf:homepage rdf:resource = "http://rdfweb.org/people/danbri/" />
  < foaf:img rdf:resource = "http://rdfweb.org/people/danbri/mugshot/danbri-small.jpeg"/>
< /foaf:Person >
```

This example encodes the information that "there is a *foaf:Person* with a foaf:name property of 'Dan Brickley' and a *foaf:mbox_sha1sum* property of 24 . . . 52e; this person stands in a *foaf:homepage* relationship to a thing called http://rdfweb.org/people/danbri/ and a *foaf:img* relationship to a thing called http://rdfweb.org/people/danbri/mugshot/danbri-small.jpeg". FOAF defines 12 classes and 51 properties. The *foaf:knows* property is used to construct basic social networks, linking to instances of *foaf:Person*.

The FOAF vocabulary is simple, which has encouraged its adoption and use, and extensible, making it suitable to a wide range of uses. As our studies have shown, more than 150 different properties have been defined for the *foaf:Person* class and nearly 500 have actually been used with instances of *foaf:Person*. One way to view this situation is that it represents undisciplined chaos and that the lack of any centralized authority or standard for terms suggests that nothing useful will come out of it. An alternate view is that communities will be able to select and use terms that are useful and those which are widely used be integrated into consensus ontologies. In this view the eventual result will be a relatively small number of widely used ontologies with mappings, as appropriate, between them. Less widely used terms, whether they are deprecated, or newly introduced, will remain on the edges.

Our investigation the most commonly used ontologies (Table I) confirms that, besides the meta-level ontologies (i.e. RDF, RDFS, DAML and OWL), one of the best populated ontology is FOAF, www.foaf-project.org/). In addition, representing personal information is also a popular theme in ontology engineering with more than a 1,000 RDF documents defining RDF terms containing the string "person"[1]. The other well populated ontologies in Table I include Dublin Core element set (DC, http://purl.org/dc/elements/1.1/), which defines document metadata properties without domain/range qualification, and RSS (RDF site summary, http://web.resource.org/rss/1.0/spec), which is "a lightweight multipurpose extensible metadata description and syndication format" for annotating web sites. FOAF provides an RDF/XML vocabulary to describe personal information (Dumbill, 2002a), including name, mailbox, homepage URL, friends, and so on. FOAF documents then induces the "web of acquaintances" (Golbeck *et al.*, 2003) and thus an implicit trust network to support such applications as knowledge outsourcing (Ding *et al.*, 2004a, b) and online communities (Dumbill, 2002b).

The advances in FOAF vocabulary and applications highlight several challenging issues. For example, how can one assemble a collection of FOAF documents to support semantic web research? What are the common patterns of connections among FOAF documents? What terms in FOAF vocabulary are the most frequently used? What is the potential of FOAF in enabling and enhancing the intelligence of web-based

Prefix	Namespace URI	Documents populated
RDF	http://www.w3.org/1999/02/22-rdf-syntax-ns#	321,108
DC	http://purl.org/dc/elements/1.1/	238,346
RSS	http://purl.org/rss/1.0/	195,018
MCVB	http://webns.net/mvcb/	110,434
FOAF	http://xmlns.com/foaf/0.1/	79,226
RDFS	http://www.w3.org/2000/01/rdf-schema#	65,486

Table I.
Best populated ontologies (generated in April 2005)

information systems? The current FOAF literature (Adamic *et al.*, 2003; Dumbill, 2002a, b, 2003; Grimnes *et al.*, 2004; Golbeck *et al.*, 2003) provides a vision and various models of how FOAF documents might be used to support web-based information system under the assumption that FOAF documents are widely available. There is still a lack of an empirical investigation on the characteristics and structure of the growing body of millions of FOAF documents. This paper presents empirical results to answer the above questions based on a large collection (over 1.5 million) of real world FOAF documents harvested from the web.

Our research on online FOAF profile documents consists of four steps: identification of FOAF documents, discovery of FOAF documents using software agents, extraction of person information, and fusion of person information based on the semantics of FOAF vocabulary. Using the statistics over this corpus, we describe the common properties and namespaces shared by the FOAF community. We hope that this analysis might help FOAF developers design and build better tools as well as inform novice FOAF users on how to create effective FOAF documents. Analyses of the social networks encoded in FOAF documents provide insight into some interesting structural patterns of the semantic web from the person perspective. The richness of profiles in FOAF documents allows us to further characterize social ties and identify friendship types.

Friendship networks connected by FOAF relationships can provide insights into features and patterns of social networks in the semantic web and advance the theories and models of social structures. Friendship networks in the physical world have been long studied in the social science. A well known example is Milgram's (1967) small-world phenomenon – the observation that everyone in the world can be reached through a short chain of social acquaintances. The concept gives rise to the famous phrase six degrees of separation, which has recently been applied to SNA in both physical and virtual environments (Xu and Chen, 2003; Adamic *et al.*, 2003). Social relationships have been derived from the contextual information or domain knowledge, e.g. co-citation relationship (Chen, 1999), indirectly using data mining techniques. In addition to social networks, the collection of FOAF documents can serve as valuable resource for semantic web research in the development and testing of trust models as well as trust propagation models (Ding *et al.*, 2004a, b).

As the first study along this line, this paper reflects the state of FOAF usage and identifies any potential problems to guide the future practice. It further contributes to the stabilization of individual terms in FOAF vocabulary. Using people as the bridge, FOAF can potentially link most of other kinds of things we describe in the web, including documents they co-authored, research interest they shared, photos they shot together, and so on. Based on relationships represented in FOAF, we can identify online communities in a research area and even discover existing communities and the emergence of new communities. As the semantic web evolves, there will be opportunities to study social dynamics and apply the findings in this study to support semantic web applications.

The remainder of this paper is organized as follows. Section 2 presents a review of the literature concerning FOAF vocabulary and SNA. Section 3 introduces a novel approach to building FOAF documents collection and analyzing the structure of friendship networks in the semantic web. Section 4 uses descriptive statistics and SNA to present findings on components of FOAF documents and structural relationships

among person profiles. Section 5 concludes with a discussion the findings of this study and their implications to the semantic web research and practice.

2. Background

2.1 FOAF document

The most important component of a FOAF document is the *FOAF vocabulary*, which is identified by the namespace URI http://xmlns.com/foaf/0.1/. The FOAF vocabulary defines both classes (e.g. *foaf:Agent*, *foaf:Person*, and *foaf:Document*) and properties (e.g. *foaf:name*, *foaf:knows*, *foaf:interests*, and *foaf:mbox*) grounded in RDF semantics. In contrast to a fixed standard, the FOAF vocabulary is managed in an open source manner, i.e. it is not stable and is open for extension[2]. Therefore, inconsistent FOAF vocabulary usage is expected across different FOAF documents.

The practical significance of FOAF to information creators and consumers can be illustrated with a variety of applications (Dumbill, 2002a, 2003), which are summarized as follows. To information publishers, FOAF is useful by

- Managing communities by offering a basic expression for community membership. Many communities have proliferated on the web, ranging from companies through professional organizations to social groups.
- Expressing identity by allowing unique user IDs across applications and services without compromising privacy. For example, the *foaf:mbox_sha1sum* property is the ASCII-encoded SHA1 hash of a mailbox URI (e.g. mailto:finin@umbc.edu). To ensure privacy, the encoding is a one-way mapping and cannot be trivially reverse-engineered.
- Indicating authorship. FOAF tools use digital signatures to associate an e-mail address with a document. Specifically, OpenPGP is used, along with the namespace http://xmlns.com/wot/0.1/ to denote concepts forming a "web of trust". This associates a signature with the document itself and specifies a signature for the linked document as part of an *rdfs:seeAlso* link. Thus, authorship information can be expressed both inside and outside of the concerned documents.

FOAF supports information consumers by:

- Allowing provenance tracking and accountability (Dumbill, 2003). On the web, the source of information is just as important as the information itself in judging its credibility. Provenance tracking RDF tools can tell where and when a piece of information is obtained. A practice common to the FOAF community is to attach the source URI to each RDF statement.
- Providing assistance to new entrants in a community. For example, people unfamiliar with a community can learn the structure and authority of a research area from the community's FOAF files.
- Locating people with common interests. Users tend to have interests and values similar to those they desire in others (Adamic *et al.*, 2003). Peer-to-peer relationships are an essential ingredient to collaboration, which is the driving force of online communities.

- Augmenting e-mail filtering by prioritizing mail from trustable colleagues. Using the degree of trust derived from FOAF files, people can prioritize incoming e-mail and thus filter out those with low trust values.

2.2 Social networks on the web

A social network consists of people or groups connected by a set of social relationships, such as friendship, co-working or information exchange (Garton *et al.*, 1997). Determining structural properties of virtual communities is the most straightforward application of SNA. The underlying physical social network can be reflected in an online community. For example, Club Nexus (Adamic *et al.*, 2003) is an online community serving over 2,000 Stanford undergraduate and graduate students. Students can use Club Nexus to send e-mail and invitations to events, post events, buy and sell goods, search and connect to people with similar interests, etc. Statistical analyses revealed that personalities and preferences of users mostly align with each other.

In addition to member relationship in online communities, SNA has been applied to many other types of social networks. For example, Xu and Chen (2003) created, analyzed and visualized a network of known criminals and their relationships. Their analysis identifies various groups and subgroups, key individuals, and links between groups. Centrality can be detected using graph properties including degree (the number of direct links), betweenness (geodesics passing through), and closeness (sum of geodesics). Each of these indices is evidence for different individual roles: a high degree suggests leadership and high betweenness indicates a "gatekeeper". This increased understanding enables law enforcement officers to target specific criminals, to disrupt criminal organizations, and to achieve higher rates of conviction.

Chen (1999) describes the development and application of visualization techniques allowing users to access and explore information in a digital library effectively and intuitively based on co-citation relationships. Salient semantic structures and citation patterns are extracted from several document collections using latent semantic indexing and pathfinder network scaling. Author co-citation patterns are visualized through a number of author co-citation maps highlighting important research areas in the field. This approach provides a means of transcending the boundaries of collections of documents and visualizing more profound patterns in terms of semantic structures and co-citation networks.

Link structure analyses and graph-theory have been applied to crawling the web for virtual communities. The FOAF project takes the social networking aspect of the web still further (Dumbill, 2002a, b), allowing the information collected to be aggregated, integrated and fused.

3. Discovering FOAF information on the web

By running the SwoogleBot (Ding *et al.*, 2004a, b, 2005) semantic web crawler in conjunction with an agent that understands FOAF vocabulary, we collected 49,750 RDF documents containing 207,413 instances of *foaf:Person* during the first three months of 2005. We intentionally limited the dataset by collecting at most 50,000 documents from any single web site and no documents from several large blog sites (e.g. www.livejournal.com).

3.1 Provenance of the data

Table II lists the five community web sites with the most number of FOAF documents. We identify several different contests in which this information is used: to describe blog authors, to describe virtual community members, or to annotate photographs.

Although community web sites have contributed large numbers of FOAF instances, their regular structure also overwhelms the variety of vocabulary and structure introduced by people who construct and self-publish FOAF profiles. We adopted a simple heuristic applied to URLs, to recognize those from community web sites. If there are a large number of URLs from a given site that differ only in a single URL argument, we classify them as automatically generated. Table III shows some extracted URL pattern[3] for community web sites.

Using this heuristic classification, we found 2,233 non-community web sites (out of 18,201) contributing 4,156 FOAF documents. We further partitioned the dataset (*GALL*) into seven subsets:

(1) Groups *G*1-*G*5 for five individual web sites contributing over 3,000 URLs:
   - *G*1 (www.wasab.dk, 4,910 urls) and *G*3 (www.kwark.org, 3,400 urls) are personal web sites mainly for annotating photos;
   - *G*2 (blog.livedoor.jp, 4266 urls), *G*4 (blogs.dion.ne.jp, 3,118 urls) are Japanese community web sites; and
   - *G*5 ((username).cocolog-nifty.com, 3,108 urls) is a Japanese blog web sites.

(2) Group *GC* contains urls from web sites being identified as community.

(3) Group *GNC* contains all urls from non-community web sites.

Host	Context	FOAF dataset	Swoogle discovered	Google site estimation
www.livejournal.com	Blog	Avoid	46,661	5,370,000
www.tribe.net	Community	Avoid	23,518	2,920,000
blog.livedoor.jp	Blog	4,266	10,120	119,000
www.greatestjournal.com	Blog	Avoid	10,097	282,000
www.wasab.dk	Annotation	4,910	8,434	73,700

Table II. Community web sites

URL pattern	Example match (amount of matches)
Same host and path, different query	www.boards.ie/network/foaf.php = ?[QUERY](2490)
Same host and path, different query	www.boards.ie/network/foaf.php = ?[QUERY](2490)
Same host, no query, path differs in one segment	http://journal.bad.lv/users/[USERNAME]/data/foaf(2548) http://blog.livedoor.jp/[USERNAME]/foaf.rdf(4242) http://swordfish.rdfweb.org/photos/genfiles/ilrt/[FILENAME](266)
Same path, no query, host differs in first segment	http://[USERNAME].cocolog-nifty.com/foaf.rdf(3108)

Table III. URL patterns for community web sites

3.2 Properties of foaf:Person

Since RDF does not have a mechanism of requiring properties for an instance, instances of foaf:Person may come with various kinds and amounts of information. We observed that only 16 properties with the domain *foaf:Person* have been defined in the original FOAF ontology and 140 more have been proposed by other ontologies according to *Swoogle ontology dictionary*. In order to evaluate their utility in practice, we collected statistics about the properties being used to describe instances of *foaf:Person*. We found 546 distinct properties used for at least one person instance, as shown in Table IV. Only 34 properties were used by more than 1 percent of the FOAF documents. The remaining properties were rarely used FOAF terms (e.g. foaf:yahooChatID), misspelled terms (e.g. foaf:firstname) or relatively new and experimental terms (e.g. foaf:mailbox and http://purl.org/vocab/relationship/spouseof).

Figure 1 shows the 15 most frequently used terms in FOAF dataset and the percentage of the documents which use each. We associate two types of property usage with two context:

(1) document usage reflects the preference of the authors' own personal information; and

(2) instance usage reflects the preference of publishing the referred persons' information.

Hence, we may find that *name*, *mbox_sha1sum*, are *rdfs:seeAlso* are preferred to describe a link to an author's friends.

We also observed the impact of community web sites in property usage as shown in Figure 2. The statistics show that:

(1) Community web sites usually make *mbox_sha1sum*, *weblog* and *nick* mandatory to all their users' profiles, and they may miss some properties, e.g. depiction for $G1$, homepage for $G2$ and surname for $G3$.

Property (in URIref form)	Document usage	percent	Instance usage	percent
http://xmlns.com/foaf/0.1/mbox_sha1sum	43,561	87.56	114,981	55.44
http://xmlns.com/foaf/0.1/name	34,951	70.25	121,498	58.58
http://xmlns.com/foaf/0.1/nick	33,584	67.51	88,217	42.53
http://xmlns.com/foaf/0.1/weblog	27,575	55.43	70,620	34.05
http://xmlns.com/foaf/0.1/homepage	18,712	37.61	56,398	27.19
http://www.w3.org/2000/01/rdf-schema#seeAlso	18,588	37.36	102,589	49.46
http://xmlns.com/foaf/0.1/knows	13,972	28.08	14,686	7.08
http://xmlns.com/foaf/0.1/depiction	11,340	22.79	12,161	5.86
http://purl.org/vocab/bio/0.1/olb	9,318	18.73	9,320	4.49
http://xmlns.com/foaf/0.1/img	8,706	17.50	8,866	4.27
http://xmlns.com/foaf/0.1/surname	6,576	13.22	9,538	4.60
http://xmlns.com/foaf/0.1/givenname	6,530	13.13	8,162	3.94
http://xmlns.com/foaf/0.1/mbox	5,327	10.71	8,463	4.08
http://xmlns.com/foaf/0.1/firstName	4,051	8.14	6,019	2.90
http://xmlns.com/foaf/0.1/page	3,795	7.63	3,851	1.86

Table IV.
Property usage in FOAF dataset

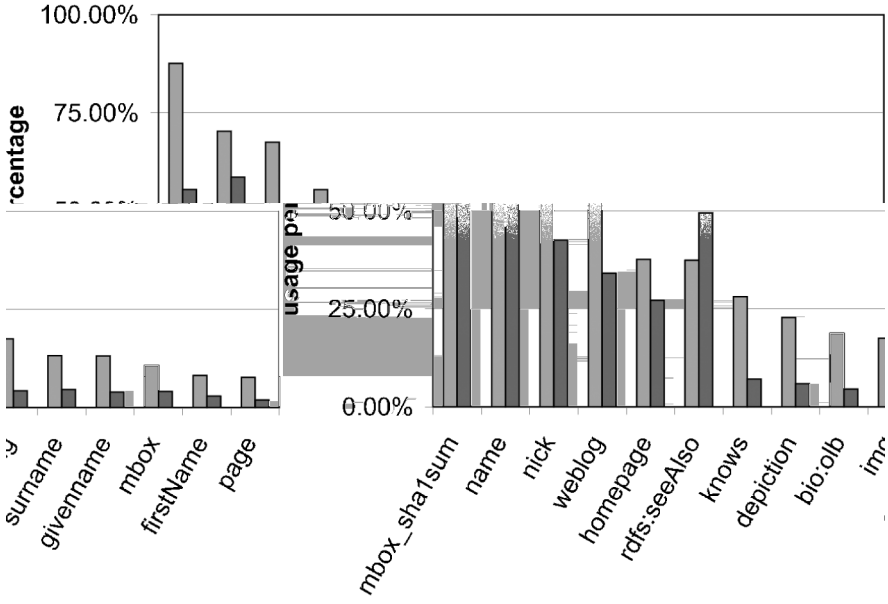


Figure 1.
Document/instance usage of best used properties

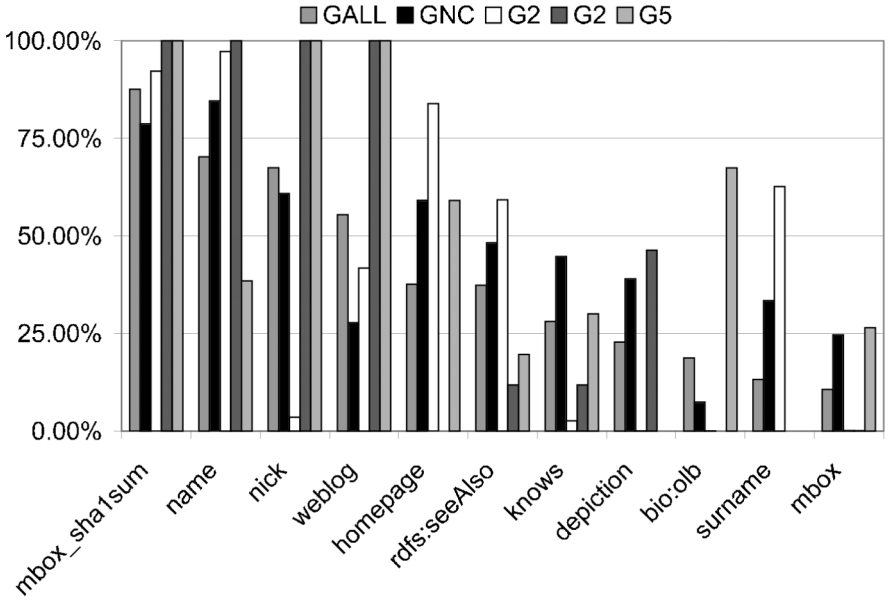


Figure 2.
Document usage of best used properties in different groups

(2) Non-community web site authors prefer name to *mbox_sha1sum*, and they usually publish their homepage, mbox, depiction (personal photographs), first name, surname, and friends.

(3) the vocabulary used by community web sites are limited in small size (*G*1:58, *G*2:8, *G*3:48, *G*4:8, and *G*5:10) while non-community contribute a very large vocabulary (522 properties).

These facts indicate that community web sites could skew the overall statistics of FOAF dataset through to their large amount of data; hence identifying community web sites is critical to a fair evaluation on the popularity of person property.

3.3 Creators and referred persons

All FOAF documents, whether manually or automatically generated, usually require that a person provide the data. Besides the creators' personal information, other persons' information are typically mentioned even when they have not published their own FOAF profile. For example, the e-mail of Dr Benjamin Grosof, a MIT professor, is reported by a document in our FOAF dataset even though he has not published any FOAF document himself.

We classify the person instances into two categories: the creators who input their personal profiles and maintain FOAF homepages, and the referred persons who are only mentioned by the creators. To this end, we adopt a simple heuristic: the referred persons usually have relatively small amount of triples while the creators have much more. As shown in Figure 3, we select seven as threshold since there is a sharp drop between seven and eight; and we result in 21,843 (10.53 percent) creators and 185,570 (89.47 percent). Another heuristic to identify the creator is to find the one person instance which is not the object of a *foaf:knows* relation.

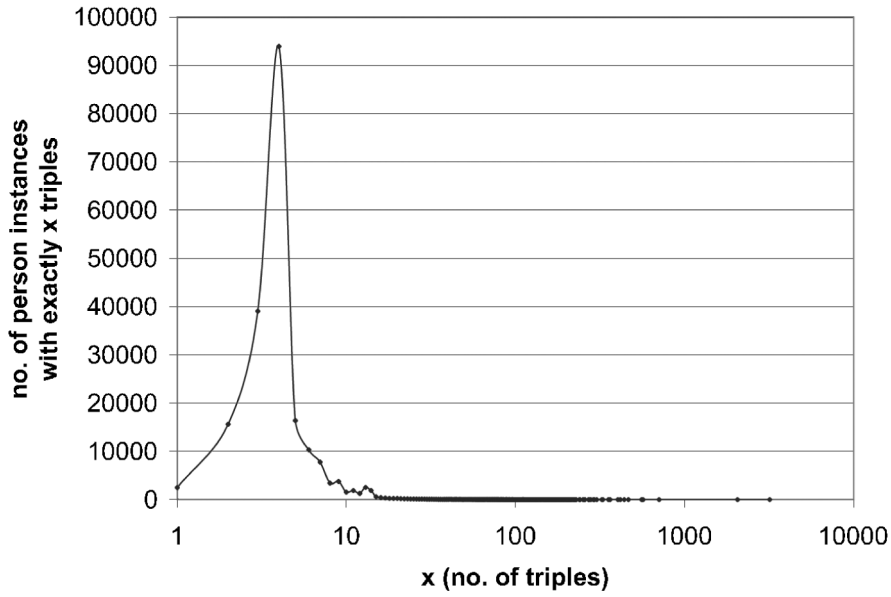


Figure 3.
Distribution of number of triples per person instance

4. Analyzing FOAF social network

We briefly outline two applications involving FOAF data. The first involves the integration and fusion of information associated with individuals. This makes good use for the fact that some FOAF properties can be declared as "inverse functional" and thus offer evidence that two individual FOAF person nodes describe the same person. The second use is to use FOAF data collected from the web as data about large-scale social networks. FOAF data describing millions of people can readily be collected from the web today offering new opportunities to explore and test social networking tools, theories and applications.

4.1 Fusing distributed personal information

One of the principles of the semantic web is that "anyone is allowed to say anything about any resource". For example, document *D*1 can make assertions about individuals introduced in document *D*2. Since FOAF is based on RDF, this allows one person to assert information about others, be they friends, acquaintances or complete strangers. Hence information about an individual may be spread across a number of FOAF documents in a collection, providing a kind of community view that mirrors the person's view in the community of people. When a person is described in more than one FOAF documents, we must fuse information from multiple sources and generate aggregated information about the person.

4.2 Person identifiers

In FOAF data, two foaf:Person instances can be identified as describing the same person in one of two ways. The first is through by URI: two non-anonymous individuals sharing the same URIref in RDF graph can be fused. The second is via assertions involving an OWL InverseFunctionalProperty. The FOAF ontology semantics defines unique identifiers of person, such as *foaf:mbox*, *foaf:mbox_sha1sum*, *foaf:homepage* and *foaf:weblog*, which are ideal clues to information fusion. In our FOAF dataset we found 644 URIrefs, 11,405 mbox_sha1sums, 6,099 homepages, 3,563 weblogs, and 757 mboxs being used as the identifiers of at least two person instances.

4.3 Fusing person information

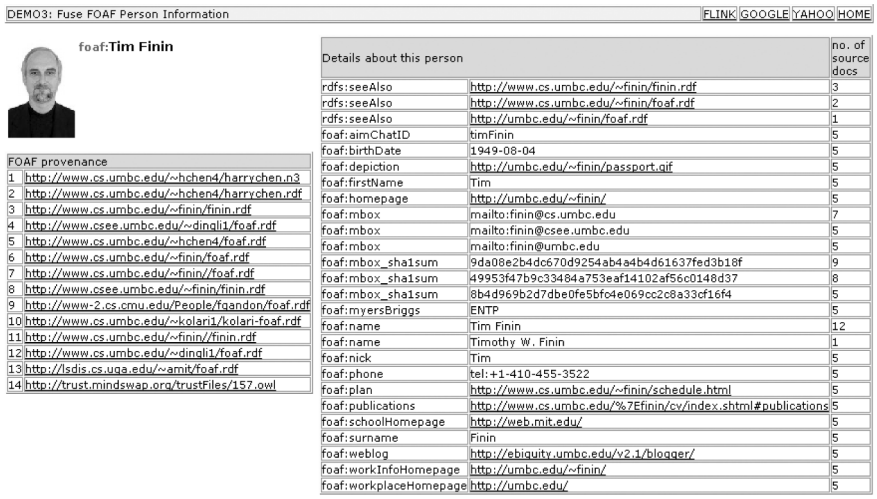
Figure 4 shows the result of fusing Dr Tim Finin's personal information from 12 sources. We found two different values of *foaf:name* from two different sources in this case:

(1) Tim Finin as stated by his FOAF profile; and

(2) Timothy W. Finin as mentioned in www-2.cs.cmu.edu/People/fgandon/foaf.rdf.

The latter is in fact the unique author identifier in DBLP[4].

Caution should be taken in merging information from multiple FOAF documents since some of the facts may be wrong and the collection of facts may contain contradictions. Small errors in FOAF documents can lead to unexpected results. For example, some FOAF documents from blog.livedoor.jp, e.g. http://blog.livedoor.jp/rusa95/foaf00756.rdf, mistakenly assign the same mbox_sha1sum to different people from 4,835 FOAF documents. We also found that Dr Jim Hendler is wrongly fused with Norman Walsh by a FOAF document in which *foaf:mbox_sha1sum* was mistakenly associated with Norman's e-mail-hash.


DEMO3: Fuse FOAF Person Information

FLINK GOOGLE YAHOO HOME

foaf:**Tim Finin**

FOAF provenance	
1	http://www.cs.umbc.edu/~hchen4/harrychen.n3
2	http://www.cs.umbc.edu/~hchen4/harrychen.rdf
3	http://www.cs.umbc.edu/~finin/finin.rdf
4	http://www.csee.umbc.edu/~dinqli1/foaf.rdf
5	http://www.cs.umbc.edu/~hchen4/foaf.rdf
6	http://www.cs.umbc.edu/~finin/foaf.rdf
7	http://www.cs.umbc.edu/~finin//foaf.rdf
8	http://www.csee.umbc.edu/~finin/finin.rdf
9	http://www-2.cs.cmu.edu/People/fgandon/foaf.rdf
10	http://www.cs.umbc.edu/~kolari1/kolari-foaf.rdf
11	http://www.cs.umbc.edu/~finin//finin.rdf
12	http://www.cs.umbc.edu/~dinqli1/foaf.rdf
13	http://lsdis.cs.uga.edu/~amit/foaf.rdf
14	http://trust.mindswap.org/trustFiles/157.owl

Details about this person		no. of source docs
rdfs:seeAlso	http://www.cs.umbc.edu/~finin/finin.rdf	3
rdfs:seeAlso	http://www.cs.umbc.edu/~finin/foaf.rdf	2
rdfs:seeAlso	http://umbc.edu/~finin/foaf.rdf	1
foaf:aimChatID	timFinin	5
foaf:birthDate	1949-08-04	5
foaf:depiction	http://umbc.edu/~finin/passport.gif	5
foaf:firstName	Tim	5
foaf:homepage	http://umbc.edu/~finin/	5
foaf:mbox	mailto:finin@cs.umbc.edu	7
foaf:mbox	mailto:finin@csee.umbc.edu	5
foaf:mbox	mailto:finin@umbc.edu	5
foaf:mbox_sha1sum	9da08e2b4dc670d9254ab4a4b4d61637fed3b18f	9
foaf:mbox_sha1sum	49953f47b9c33484a753eaf14102af56c0148d37	8
foaf:mbox_sha1sum	8b4d969b2d7dbe0fe5bfc4e069cc2c8a33cf16f4	5
foaf:myersBriggs	ENTP	5
foaf:name	Tim Finin	12
foaf:name	Timothy W. Finin	1
foaf:nick	Tim	5
foaf:phone	tel:+1-410-455-3522	5
foaf:plan	http://www.cs.umbc.edu/~finin/schedule.html	5
foaf:publications	http://www.cs.umbc.edu/%7Efinin/cv/index.shtml#publications	5
foaf:schoolHomepage	http://web.mit.edu/	5
foaf:surname	Finin	5
foaf:weblog	http://ebiquity.umbc.edu/v2.1/blogger/	5
foaf:workInfoHomepage	http://umbc.edu/~finin/	5
foaf:workplaceHomepage	http://umbc.edu/	5

Figure 4.
Fusing Dr Tim Finin's personal information

4.4 Social network analysis

A collection of distributed FOAF documents may constitute a social network. The *foaf:knows* relation can link one individual of *foaf:Person* to another. The FOAF dataset contains 131,314 triples and produces 109,470 *foaf:knows* relations among 49,861 instances in FOAF dataset after fusing person. We focused on a smaller portion of that big social network – the emerging social networks in the distributed semantic web, which is different than those emerged from a centralized community web site. Therefore, our analyses only concerned FOAF documents from non-blog web sites. We found many instances followed Zipf's (1932) distributions, so all the figures of distribution were plotted on log-log scale.

4.4.1 Social network from dataset GNC. We selected about 4,156 FOAF documents containing 32,727 FOAF person instances before fusing persons. After fusing, we obtain a social network SN_{GNC} with 15,630 foaf:knows relations among 26,788 persons. Only 2,799 (10 percent) persons are really fused from at least two original person instances. People fused from many sources could be either social authorities, who are known by many people, or semantic web experts (blogger as well), who maintain a fairly large amount of FOAF documents. The top ten people are listed as the following with the amount of original instances they fused from.

- Social authorities, who are known by many people. For example, Danny Ayers (386), Dan Brickley (199), Libby Miller (133), Edd Dumbill (76), Morten Frederiksen (48), Charles McCathieNevile (39), Dan Connolly (35), Marc Canter (33), Peter Mika (32).
- Semantic web experts, who are usually an active blogger and maintain a "personal" web site with large amount of FOAF documents. For example, Christoph Görn (719), Ian Davis (360), Christopher Schmidt(196), Jim Ley(124), Vincent Tabard (71), Masahide Kanzaki (60).
- Figure in photo, who have been mentioned by a lot of photo annotation. ONO Hiroki (134), Libby Miller (133), Gregory Todd Williams (61).

4.4.2 Patterns of degree. Degree analysis is an important tool in SNA. Our analyses were based on 15,630 "knows" links within GNC. Figures 5 and 6 show the distributions of in-degrees and out-degrees, respectively. It is shown that only a few fused persons have more than one in-degrees or out-degrees. In fact, among the 26,788 fused persons, only 11.62 percent of them have both in-links and out-links, and 78.11 percent of them have only one in-link. All this statistics indicates the sparseness of the SN_{GNC}.

4.4.3 Patterns of connected components. There are 842 components in SN_{GNC} with average size 16. The distribution of component size is highly skewed as shown in figure Error! Reference source not found: there is one very large component with 7,111 fused individuals and the second with only 549 (less than 10 percent of the size of the first). We note that the large component was fused due to errors in the FOAF documents, which mistakenly assigned the same *foaf:mbox_sha1sum* to many different individuals (Figure 7).

The inherent nature of FOAF publishing makes the star-shaped component shown in Figure 8[5] common. It typically arises when an individual publishes a FOAF document describing a set of people with whom they have a *foaf:knows* relationship.

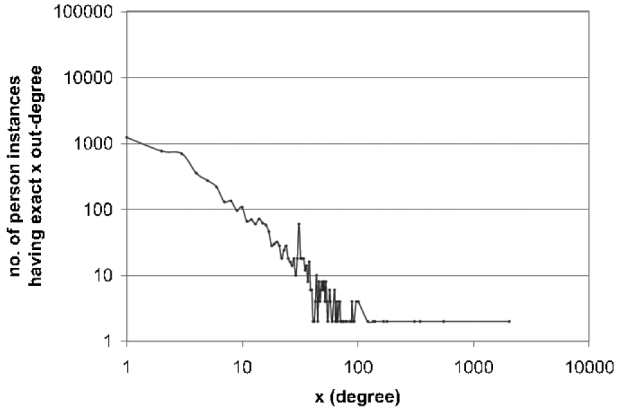


Figure 5.
Out-degree distribution per group

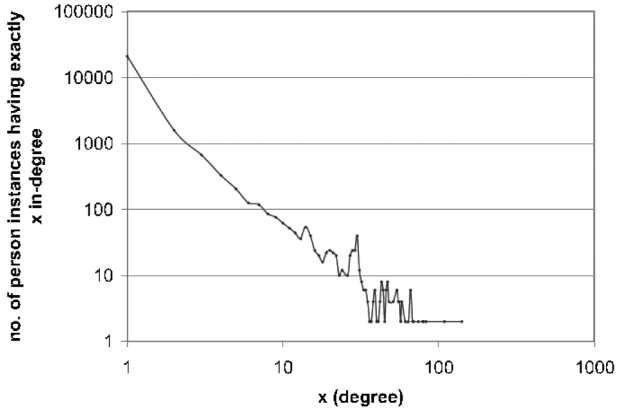


Figure 6.
In-degree distribution per group

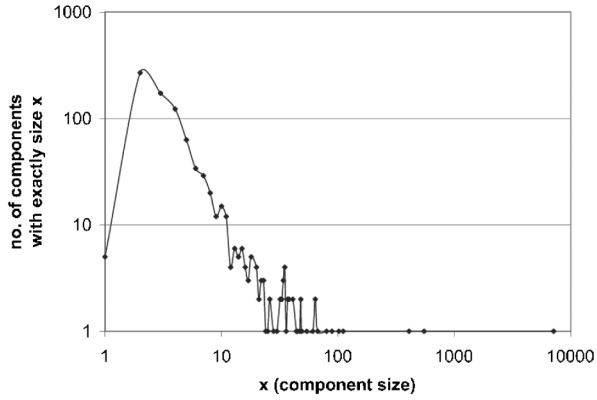


Figure 7.
Distribution of component size

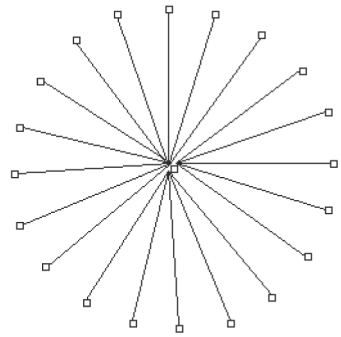
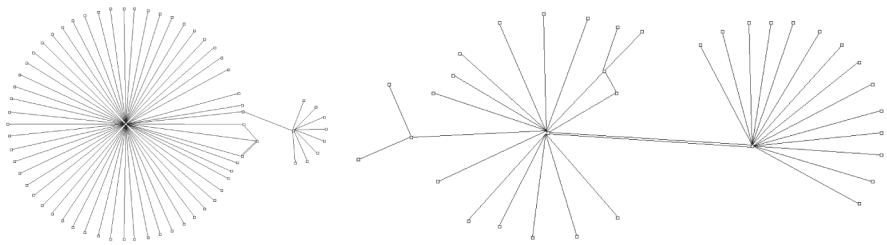


Figure 8.
Star-shaped components are common in naturally occurring FOAF profiles

As more people publish FOAF profiles, the star configurations of the early adopters may get their influence spread in bud mode as shown in Figure 9(a) or hook up with each other though bi-directional bridges, as shown in Figure 9(b).

The second largest component in SN_{GNC}, as shown in Figure 10 with 546 nodes and 771 directed edges, turns out to be a proof of the above social network growth models. It features several hubs with very high out-degree, plus several other nodes with extremely high betweenness value staying between those hub nodes.



(a) expanding the next level

(b) two social networks being bridged

Figure 9.
Component growth models

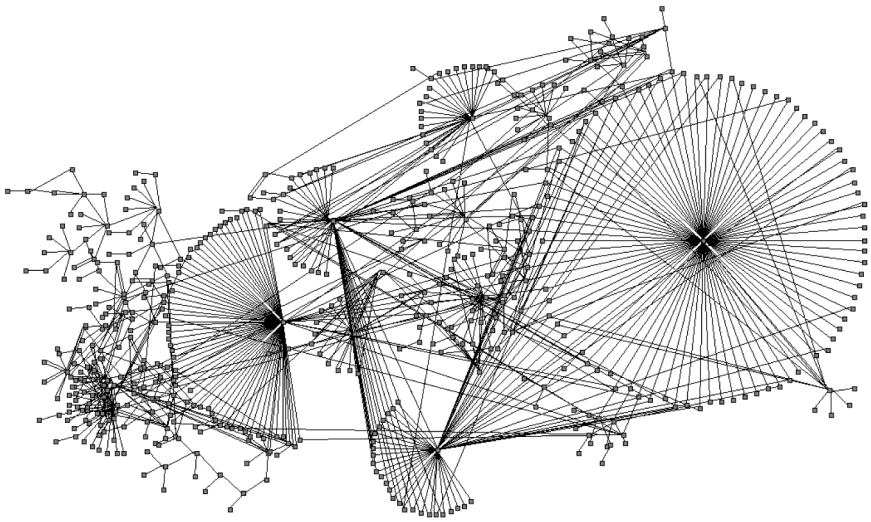


Figure 10.
The second largest component in SN_{GNC}

5. Conclusions

The semantic web offers an ambitious vision of an internet populated with intelligent agents and services able to exchange information, tasks and knowledge using simple protocols coupled with a rich knowledge representation language. Exploring the roadmap leading toward this vision will take some time. The semantic web languages RDF and OWL are a promising beginning. One of the first wide-spread applications of RDF is the representation of social networks – individuals, their properties and the relationships among them. The current interest in social networks and the immediate applications to online virtual communities have made the FOAF ontology the widely used on the web at this writing. Studying how FOAF is being used provides a good test case for the larger questions and issues involving the adoption of semantic web concepts and technologies.

We presented a novel perspective of the semantic web by linking machine-readable descriptions of people, i.e. FOAF documents, with published personal relationships. This complements the ontology-based view of the semantic web. We also proposed a heuristic approach to identifying and discovering FOAF documents from the web and extracting information about people from these FOAF documents. This approach provides a means of transcending the boundaries of individual FOAF documents, fusing information about a person from multiple documents. The analysis of FOAF network pattern also lent itself to unique social network structures in the semantic web.

FOAF networks provide a snapshot of the FOAF user community encoded in the constituent *foaf:knows* relations. More importantly, connection patterns among FOAF documents offer a persons orientation to the conventional web of HTML documents. The visualization of highly connected FOAF networks is informative and revealing. As the number of FOAF users grows, the approach presented in this paper can be used to discover existing and emerging online communities.

Notes

1. This is reported by our Swoogle (http://swoogle.umbc.edu), a RDF crawling and indexing engine [Error! Reference source not found.].
2. The lastest FOAF specification only lists one stable term – "homepage" and leaves many others in "testing" or "unstable" stages.
3. The syntax of URL is based on RFC 2396, and we follow the convention that a URL has four components " < scheme > :// < authority > < path > ? < query > " and we concentrate on the host part of an authority.
4. www.informatik.uni-trier.de/ley/db/
5. Figure Error! Reference source not found. – 11 were generated by the "Otter" network visualization tool (Error! Reference source not found.).

References

Adamic, L.A., Buyukkokten, O. and Adar, E. (2003), "A social network caught in the web", *First Monday*, Vol. 8 No. 6.

Chen, C. (1999), "Visualising semantic spaces and author co-citation networks in digital libraries", *Inf. Process Manage*, Vol. 35 No. 3, pp. 401-20.

Ding, L., Zhou, L., Finin, T. and Joshi, A. (2005), "How the semantic web is being used: an analysis of FOAF", *Proceedings of the 38th International Conference on System Sciences*, Hawaii, January.

Ding, L., Kolari, P., Ganjugunte, S., Finin, T. and Joshi, A. (2004a), "Modeling and evaluating trust network inference", paper presented at Seventh International Workshop on Trust in Agent Societies, The Third International Joint Conference on Autonomous Agents and Multi Agent Systems, New York, NY, July.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V.C. and Sachs, J. (2004b), "Swoogle: a search and metadata engine for the semantic web", *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*.

Dumbill, E. (2002a), "Finding friends with XML and RDF", IBM's XML Watch, available at: www-106.ibm.com/developerworks/xml/library/x-foaf.html

Dumbill, E. (2002b), "Support online communities with foaf: how the friend-of-a-friend vocabulary addresses issues of accountability and privacy", IBM's XML Watch, available at: www-106.ibm.com/developerworks/xml/library/x-foaf2.html

Dumbill, E. (2003), "Tracking provenance of RDF data", IBM's XML Watch, available at: www-106.ibm.com/developerworks/xml/library/x-rdfprov.html

Garton, L., Haythornthwaite, C. and Haythornthwaite, C. (1997), "Studying online social networks", *Journal of Computer-Mediated Communication*, Vol. 3.

Golbeck, J., Parsia, B. and Hendler, J. (2003), "Trust networks on the semantic web", *Proceedings of Cooperative Intelligent Agents*.

Grimnes, G.A., Edwards, P. and Preece, A. (2004), "Learning meta-descriptions of the FOAF network", *Proceedings of International semantic web Conference*.

Hayes, P. (Ed.) (2004), "Rdf semantics", w3c recommendation, available at: www.w3.org/TR/2004/REC-rdf-mt-20040210/(accessed: 10 February 2004).

Klyne, G. and Carroll, J.J. (Eds) (2004), "Resource description framework (RDF): concepts and abstract syntax", available at: www.w3.org/TR/2004/REC-rdf-concepts-20040210/

Milgram, S. (1967), "The small world problem", *Psychology Today*, Vol. 1 No. 1, pp. 60-7.

Schreiber, G. and Dean, M. (2004), "Owl web ontology language reference", February, available at: www.w3.org/TR/2004/REC-owl-ref-20040210/

Xu, J. and Chen, H. (2003), "Untangling criminal networks: a case study. Intelligence and security informatics", paper presented at First NSF/NIJ Symposium (2665), pp. 232-48.

Zipf, G.K. (1932), *Selected Studies of the Principle of Relative Frequency in Language*, Harvard University Press, Cambridge, MA.

Further reading

Ding, L., Zhou, L. and Finin, T. (2003), "Trust based knowledge outsourcing for semantic web agents", *Proceedings of the IEEE/WIC International Conference on Web Intelligence*.

Huffaker, B. (1998), Otter: a general-purpose network visualization tool.

Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A. (1999), "Trawling the web for emerging cyber-communities", *Proceedings of the Eighth International Conference on World Wide Web*, Elsevier North-Holland, Inc., Amsterdam, pp. 1481-93.