Available online at www.sciencedirect.com
SCIENCE DIRECT®
Web Semantics: Science, Services and Agents on the World Wide Web 3 (2005) 211–223
Web Semantics
www.elsevier.com/locate/websem

Flink: Semantic Web technology for the extraction and analysis of social networks

Peter Mika*

Department of Computer Science, Vrije Universiteit Amsterdam (VUA), De Boelelaan 1081, 1081HV Amsterdam, The Netherlands

Received 24 May 2005; accepted 25 May 2005

Abstract

We present the Flink system for the extraction, aggregation and visualization of online social networks. Flink employs semantic technology for reasoning with personal information extracted from a number of electronic information sources including web pages, emails, publication archives and FOAF profiles. The acquired knowledge is used for the purposes of social network analysis and for generating a web-based presentation of the community. We demonstrate our novel method to social science based on electronic data using the example of the Semantic Web research community.

Keywords: Semantic Web; Social networks; Ontology extraction; Social ontology

1. Introduction

The possibility to publish and gather personal information (such as the interests, works and opinions of our friends and colleagues) has been a major factor in the success of the web from the beginning. Remarkably, it was only in the year 2003 that the web has become an active space of socialization for the majority of users. That year has seen the rapid emergence of a new breed of web sites, collectively referred to as social networking services (SNS). The first-mover Friendster[1] attracted over 5 million registered users in the span of a few months [13], which was followed by Google and Microsoft starting or announcing similar services.

Although these sites feature much of the same content that appear on personal web pages, they provide a central point of access and bring structure in the process of personal information sharing and online socialization. Following registration, these sites allow users to post a profile with basic information, to invite others to register and to link to the profiles of their friends. The system also makes it possible to visualize and browse the resulting network in order to discover friends in common, friends thought to be lost or potential new friendships based on shared interests. (Thematic sites cater to more specific goals, such as establishing a business contact or finding a romantic relationship.)

* Tel.: +31 20 5987753; fax: +31 20 5987653.
E-mail address: pmika@cs.vu.nl
URL: http://www.cs.vu.nl/pmika
[1] http://www.friendster.com

1570-8268/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2005.05.006

The latest breed of social networking services combine social networks with the sharing of content such as bookmarks, documents, photos, reviews. The idea of network-based knowledge sharing is based on the sociological theory that social interaction creates similarity and vice versa, interaction creates similarity (friends are likely to have acquired or develop similar interests). Lately, the notion of ratings and social networks-based trust are also investigated as a filtering mechanism in loosely controlled environments.

Despite their early popularity, users have later discovered a number of drawbacks to centralized social networking services. First, the information is under the control of the database owner who has an interest in keeping the information bound to the site. The profiles stored in these systems cannot be exported in machine processable formats, and therefore the data cannot be transferred from one system to the next. (As a result, the data needs to be maintained separately at different services.) Second, centralized systems do not allow users to control the information they provide on their own terms. Although Friendster follow-ups offer several levels of sharing (e.g. public information versus only for friends), users often still find out the hard way that their information was used in ways that were not intended.

These problems have been addressed with the use of Semantic Web technology. The friend-of-a-friend (FOAF) project[2] is a first attempt at a formal, machine processable representation of user profiles and friendship networks. Unlike with Friendster and similar sites, FOAF profiles are created and controlled by the individual user and shared in a distributed fashion.[3] Much like the way web pages are linked to each other by anchors, these profiles link to the profiles of friends by using the *rdfs:seeAlso* relation, creating the so-called FOAF-web.

The alert reader may note that for the purposes described above, namely providing a structured representation of user profiles, the use of XML technologies would have sufficed. In fact, the real value of FOAF is that it represents an agreement on key terms and that it is described in a semantic format (namely, OWL full). These properties make FOAF the ideal basis for the semantic integration of personal information extracted from heterogeneous knowledge sources.[4]

Flink, our system to be introduced is the first to our knowledge that exploits FOAF for the purposes of *social intelligence*. By social intelligence, we mean the semantics-based integration and analysis of social knowledge extracted from electronic sources under diverse ownership or control. In our case, these sources are largely the natural byproducts of the daily work of a community: HTML pages on the web about people and events, emails and publications. From these sources, Flink extracts knowledge about the social networks of the community and consolidates what is learned using a common semantic representation, namely the FOAF ontology.

The raison d'etre of Flink can be summarized in three points. First, Flink is a demonstration of the latest Semantic Web technology (and as such a recipient of the Semantic Web Challenge Award of 2004). In this respect, Flink is interesting to all those who are planning to develop systems using Semantic Web technology for similar or different purposes. Second, Flink is intended as a portal for anyone who is interested to learn about the work of the Semantic Web community, as represented by the profiles, emails, publications and statistics. Hopefully Flink will also contribute to bootstrapping the nascent FOAF-web by allowing the export of the knowledge in FOAF format. This can be taken by the researchers as a starting point in setting up their own profiles, thereby contributing to the portal as well. Lastly, but perhaps most importantly, the data collected by Flink is used for the purposes of social network analysis, in particular learning about the nature of power and innovativeness in scientific communities.

In this paper the focus is on the first two aspects of Flink. We begin with the introduction of the system from a user perspective in Section 2. In Section 3, we describe the architecture of Flink in detail and discuss the lessons that have been learned while developing its components. We briefly introduce the idea of network analysis using Flink in Section 4. Related and future work are discussed in the last two sections of this paper.

[2] http://www.foaf-project.org.

[3] FOAF profiles are typically posted on the personal website of the user and linked from the user's homepage with the HTML LINK tag.

[4] While FOAF carries a necessary level of commitment, the maintainers of ontology are also careful not to overly restrict the interpretation of the ontology in order to keep its wide appeal to different communities and usage scenarios.

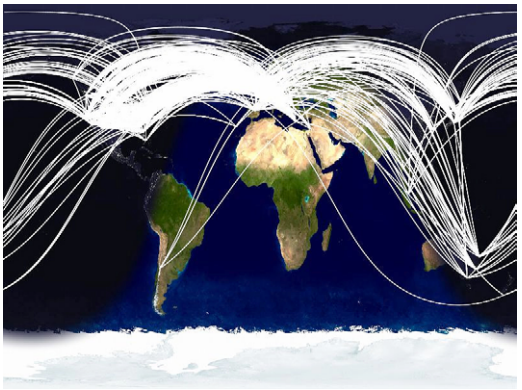


Fig. 1. Semantic Web researchers and their connections across the globe.

2. Flink: a who is who of the Semantic Web

Flink is a presentation of the professional work and social connectivity of Semantic Web researchers. For our purposes, we have defined this community as those researchers who have submitted publications or held an organizing role at any of the past International Semantic Web Conferences (ISWC02, ISWC03, ISWC04) or the Semantic Web Working Symposium (SWWS01).[5] This means a community of 608 researchers from both academia and industry, covering much of the United States, Europe and to lesser degree Japan and Australia (see Fig. 1).

Flink takes a network perspective on the Semantic Web community, which means that the navigation of the website is organized around the social network of researchers. Once the user has selected a starting point for the navigation, the system returns a summary page of the selected researcher, which includes profile information as well as links to other researchers that the given person might know. The immediate neighbourhood of the social network (the ego-network of the researcher) is also visualized in a graphical form (see Fig. 2).

The profile information and the social network is based on the analysis of webpages, emails, publications and self-created profiles. (See the following section for the technical details.) The displayed profile information includes the name, email, homepage, image, affiliation and geographic location of the researcher, as well as his interests, participation at Semantic Web related conferences, emails sent to public mailing lists and publications written on the topic of the Semantic Web. The full text of emails and publications can be accessed by following external links. At the time of writing,[6] the system contained information about 5147 publications authored by members of the community and 8185 messages sent via five Semantic Web-related mailing lists.

The navigation from a profile can also proceed by clicking on the names of co-authors, addressees or others listed as known by this researcher. In this case, a separate page shows a summary of the relationship between the two researchers, in particular the evidence that the system has collected about the existence of this relationship. This includes the weight of the link, the physical distance, friends, interests and depictions in common as well as emails sent between the researchers and publications written together.

The information about the interests of researchers is also used to generate an ontology of the Semantic Web community. The concepts of this ontology are research topics, while the associations between the topics are based on the number of researchers who have an interest in the given pair of topics (see Fig. 3). An interesting feature of this ontology is that the associations created are specific to the community of researchers whose names are used in the experiment. This means that unlike similar lightweight ontologies created from a statistical analysis of generic web content, this ontology reflects the specific conceptualizations of the community that was used in the extraction process (see the following section). Also, the ontology naturally evolves as the relationships between research topics changes (e.g. as certain fields of research move closer to each other). For a further discussion on the relation between sociability and semantics, we refer the reader to [17].

The visitor of the website can also view some basic statistics of the social network. Degree, closeness and betweenness are common measures of importance or influence in social network analysis, while the degree distribution attests to a general characteristic of

[5] An common alternative way of defining the boundary of scientific communities is to look at the authorship of representative journals (see e.g. [10]). However, the Semantic Web has a dedicated journal only since 2004 and many Semantic Web related publications appear in journals not entirely devoted to the Semantic Web.

[6] May 14, 2005.

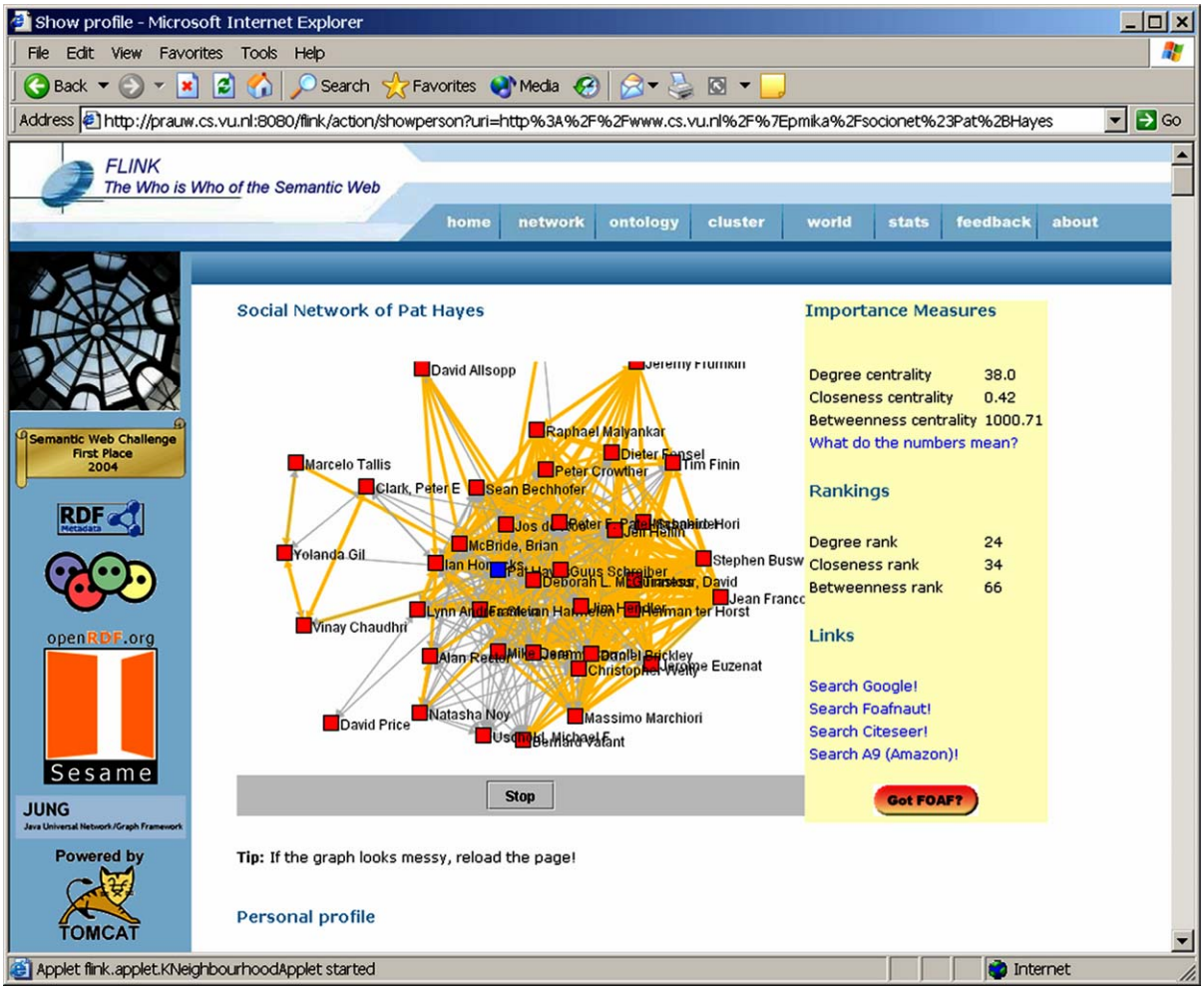


Fig. 2. The social network of a researcher.

the network itself (see Fig. 4). Geographic visualizations of the Semantic Web offer another overview of the network by showing the places where researchers are located and the connections between them (see Fig. 1).

3. System design

Similarly to the design of most Semantic Web applications, the architecture of Flink can be divided in three layers concerned with metadata acquisition, storage and visualization, respectively. Fig. 5 shows an overview of the system architecture with the three layers arranged from top to bottom. In the following, we describe the layers in the same order.

3.1. Acquisition

This layer of the system concerns the acquisition of metadata. Flink uses four different types of knowledge sources: HTML pages from the web, FOAF profiles from the Semantic Web, public collections of emails and bibliographic data. Information from the different sources is collected in different ways but all the knowledge that is learned is represented according to the same ontology (see the following section). This on-

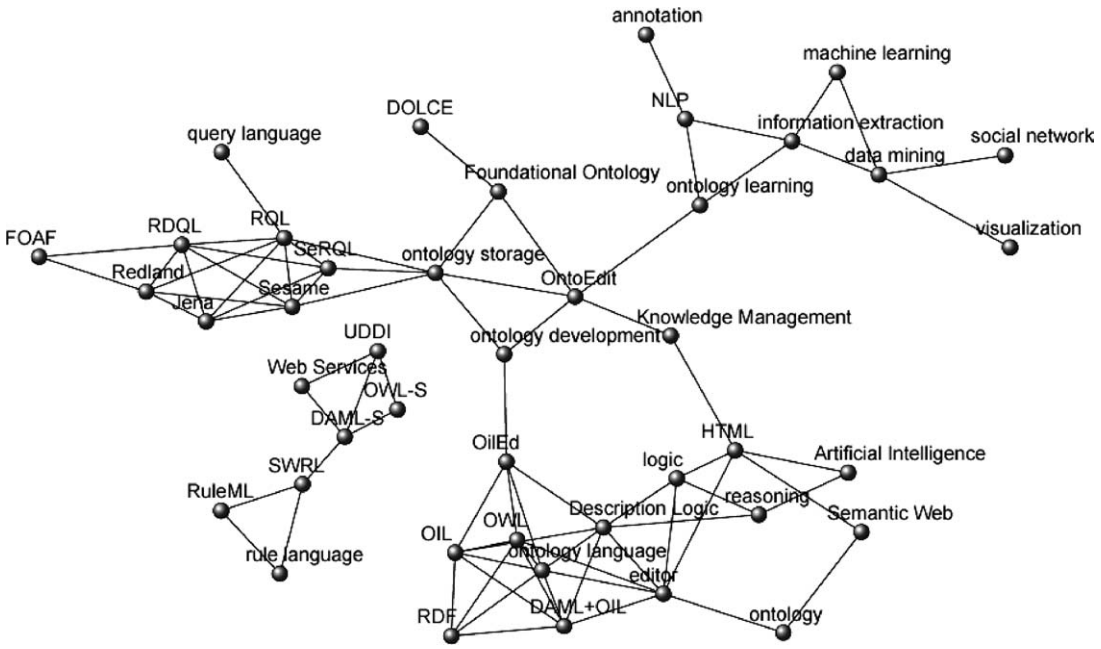


Fig. 3. The ontology of research topics.

tology includes FOAF and minimal extensions required to represent additional information.

The web mining component of Flink employs a co-occurrence analysis technique first applied to social network extraction in the work of Kautz et al. [14]. Given a set of names as input, this component of the system uses the search engine Google to obtain hit counts for the individual names as well as the co-occurrences. (The term "Semantic Web OR ontology" is added to the query for disambiguation.) The strength of association between individuals is then calculated by normalizing separately with the page counts of the individuals. The resulting value is a non-negative real number from a power-law distribution. We consider this value as evidence of a directed tie if it reaches a certain predefined threshold and the hit counts for the individuals are also above a certain minimum, in order to ensure that the support for the co-occurrence is high enough.

The web mining component also performs the additional task of finding topic interests, i.e. associating researchers with certain areas of research. Given a set of names and a list of interests (or any other kind of domain concept), the system calculates the so-called Google Mindshare for each researcher to determine whether a given person is associated with a certain interest or not. The Google Mindshare of a person with respect to an interest is simply the number of the pages where the names of the interest and the person co-occur divided by the total number of pages about the person. Note that we do not factor in the page count of the interests, since we are only interested in the expertise of the individual relative to himself.[7] The resulting measure is again a zero or positive real term with a power-law distribution. We assign the expertise to an individual if the logarithm of this value is at least one standard deviation higher than the mean of the logarithmic values. (Note that we are following here are a 'rule of thumb' in network analysis practice.)

FOAF profiles are gathered from the Semantic Web in two steps. First, an RDF crawler (a so-called scutter) is started to collect profiles from the FOAF-web.

[7] By normalizing with the hit count of the interests, the measure would assign a relatively high score—and an overly large number of interests—to individuals with many pages on the web. Since we normalize only with the page count of the person involved, we cannot compare the association strength across interests. However, this is not necessary for our purposes.

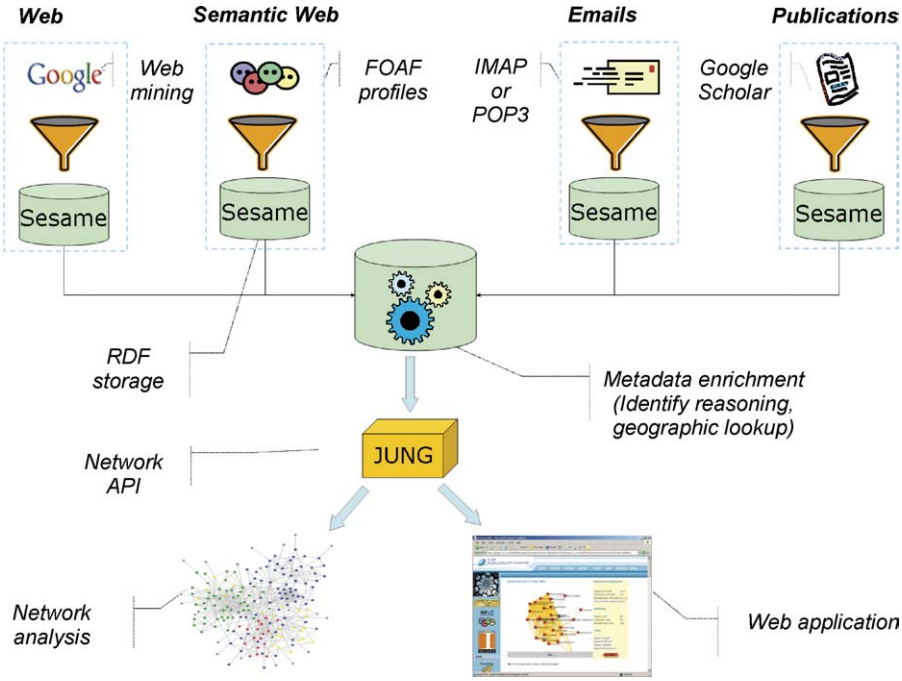


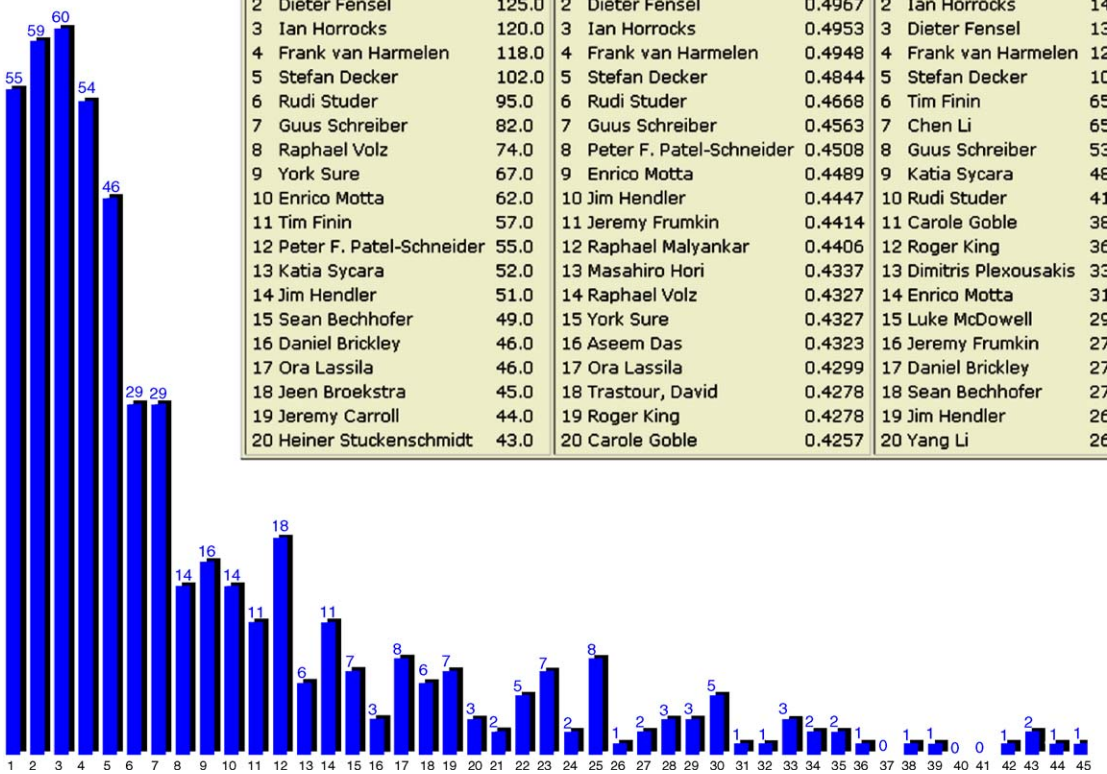
Fig. 4. The architecture of Flink from metadata acquisition (top) to the user interface (bottom).

A scutter works similar to an HTML crawler in that it traverses a distributed network by following the links (*rdfs:seeAlso* properties) from one document to the next. Our scutter is focused in that it only collects potentially relevant statements, i.e. those triples where the predicate is in the RDF, RDF-S, FOAF or WGS-84 namespace. The scutter also has a mechanism to avoid large FOAF producers that are unlikely to provide relevant data, in particular blog sites. (The overwhelming presence of these sites also make FOAF characterization difficult, see [11].) The scutter also discards documents that are simply too large, and therefore unlikely to contain a personal profile. These restrictions are necessary to limit the amount of data collected, which can easily reach millions of triples after running the scutter for only an hour. In a second step, the FOAF individuals found in the collection are matched against the profiles of the members of the target community to filter out relevant profiles from the collection. (See the following section.)

Information from emails is also processed in two steps. In this case, the first step requires that the emails are downloaded from a POP3 or IMAP store and the relevant header information is captured in an RDF format, where FOAF is used for representing information about senders and receivers of emails, in particular their name (as appears in the header) and email address. The second step is then the same as above.

Lastly, bibliographic information is collected in a single step by querying Google Scholar with the names of individuals (plus the disambiguation term). From the results, we learn the title and locations of publications as well as the year of publication and the number of citations where available.[8] This knowledge is represented in the SWRC ontology format (except for citation counts, which cannot be expressed). An alternative source of bibliographic information (used in previous versions of the system) is the Bibster peer-to-peer network [9], from which metadata can be exported directly in the SWRC ontology format.

[8] Note that it is not possible to find co-authors using Google Scholar, since it suppresses the full list of authors in cases where the list would be too long. Fortunately, this is not necessary when the list of authors is known in advance.



Indegree centrality		Closeness Centrality		Betweenness Centrality	
1 Steffen Staab	130.0	1 Steffen Staab	0.4995	1 Steffen Staab	16864.4403
2 Dieter Fensel	125.0	2 Dieter Fensel	0.4967	2 Ian Horrocks	14098.6831
3 Ian Horrocks	120.0	3 Ian Horrocks	0.4953	3 Dieter Fensel	13544.063
4 Frank van Harmelen	118.0	4 Frank van Harmelen	0.4948	4 Frank van Harmelen	12537.5246
5 Stefan Decker	102.0	5 Stefan Decker	0.4844	5 Stefan Decker	10238.5444
6 Rudi Studer	95.0	6 Rudi Studer	0.4668	6 Tim Finin	6547.5342
7 Guus Schreiber	82.0	7 Guus Schreiber	0.4563	7 Chen Li	6506.0598
8 Raphael Volz	74.0	8 Peter F. Patel-Schneider	0.4508	8 Guus Schreiber	5372.3001
9 York Sure	67.0	9 Enrico Motta	0.4489	9 Katia Sycara	4826.9771
10 Enrico Motta	62.0	10 Jim Hendler	0.4447	10 Rudi Studer	4178.6547
11 Tim Finin	57.0	11 Jeremy Frumkin	0.4414	11 Carole Goble	3882.1959
12 Peter F. Patel-Schneider	55.0	12 Raphael Malyankar	0.4406	12 Roger King	3686.1388
13 Katia Sycara	52.0	13 Masahiro Hori	0.4337	13 Dimitris Plexousakis	3337.4544
14 Jim Hendler	51.0	14 Raphael Volz	0.4327	14 Enrico Motta	3107.1791
15 Sean Bechhofer	49.0	15 York Sure	0.4327	15 Luke McDowell	2950.3047
16 Daniel Brickley	46.0	16 Aseem Das	0.4323	16 Jeremy Frumkin	2792.4088
17 Ora Lassila	46.0	17 Ora Lassila	0.4299	17 Daniel Brickley	2780.8619
18 Jeen Broekstra	45.0	18 Trastour, David	0.4278	18 Sean Bechhofer	2714.4584
19 Jeremy Carroll	44.0	19 Roger King	0.4278	19 Jim Hendler	2625.8347
20 Heiner Stuckenschmidt	43.0	20 Carole Goble	0.4257	20 Yang Li	2608.5416

Fig. 5. Simple network statistics such as the degree distribution of the network (shown on the bar chart) and the most common importance measures (shown in table) are available through the web. Other statistics can be computed by exporting the data to network analysis packages such as Pajek or UCINET.

3.2. Representation, inference and storage

This is the middle layer of our system with the primary role of storing and enhancing metadata through reasoning.

The network ties, the interest associations and other metadata are represented in RDF using terms from the FOAF vocabulary such as *foaf:knows* for relationships and *foaf:topic_interest* for research interests. (FOAF is the native format of profiles collected from the Semantic Web.) A reification-based extension of the FOAF model is necessary to represent association weights. (For a more detailed treatment of current issues in social ontology, we refer the reader to [19]).

Extensions to the FOAF model are also necessary to record the context of the statements collected.[9] Currently, this is also expressed using the RDF reification mechanism, which significantly adds to the amount of data that needs to be handled. We hope that in the future our storage facility will provide native features for context support, which would improve the efficiency of storing and querying such information. This support would be also necessary to implement efficient updates of the information.

[9] Context in our system consists of the source of a statement and the time it was collected.

The aggregated collection of RDF data is stored in a Sesame server. (For more information about the Sesame RDF storage and query facility, we refer to [4].) Note that since the model is a compatible extension of FOAF, the knowledge can be further processed from this point by any FOAF-compatible tool. An example of that is a generic component we incorporated for finding the geographical locations (latitude and longitude coordinates) of place names found in the FOAF profiles. This component invokes the ESRI Place Finder Sample Web Service, which provides geographic locations of over 3 million place names worldwide.[10] Web Service invocation is facilitated by the Apache Web Service Invocation Framework, which uses the WSDL profile of a web service to generate the code required to interact with the service.

Besides storage, inference is another major task of the middle layer. Sesame applies the RDF closure rules to the data at upload time. This feature can be extended by defining domain-specific inference rules in Sesame's custom rule language. (Note that barring a standard rule language for the Semantic Web, this remains as a practical alternative.) We use this facility to express mappings and metaknowledge, for example that co-authors of publications and senders/receivers of emails know each other in the FOAF sense.

Flink also makes use of the rule language for carrying out identity reasoning, otherwise known as smushing. Identity reasoning is required to determine the identity of instances (in this case individuals) across multiple information sources. The methods for smushing in Flink are based on name matching and object identification based on the inverse-functional properties (IFPs) of FOAF. IFP-based matching is directly axiomatized in the rule language. IFPs of the *foaf:Person* class include mailbox, mailbox checksum, homepage and several other properties. For example, if we find that instances A and B of the *foaf:Person* class have the same value for the *foaf:mbox* property, we can conclude A *owl:sameAs* B. Name matching is implemented in code and is based on the similarity of names as strings. (Differences in the last names are disallowed, however.) When matches are found, the match is again recorded using the *owl:sameAs* property.

The merging of profile information is based on the semantics of the *owl:sameAs* relation. Since Sesame has no built-in support for OWL equivalence, we axiomatize the *owl:sameAs* property using the rule language as well. The rule-based expansion of equivalence has the disadvantage that it requires the storage of the same information about all the equivalent instances. In principle, the repository could be 'cleaned' by removing all but one of the equivalent instances. However, the size of the repository is still moderate (also due to the filtering of irrelevant person instances) and the removal of statements would likely require significant additional processing.

From a scalability perspective, we are glad to note that the Sesame server offers very high performance in storing data on the scale of millions of triples, especially using native repositories. (Native storage refers to a file-system-based back-end as opposed to repositories built on top of relational databases.) Speed of upload is particularly important for the RDF crawler, which itself has a very high throughput. Unfortunately, the speed of upload drops significantly when custom rules need to be evaluated.

While the speed of uploads is important to keep up with other components that are producing data, the time required for resolving queries determines the responsiveness of the user interface. At the moment query optimization is still a significant challenge for the server. In many cases, the developer himself can improve the performance of a query by rewriting it manually, e.g. by reordering the terms or breaking the query in two. The trade-off between executing many small queries versus executing a single large query also requires the careful judgement of the developer. The trade-off is in terms of memory footprint versus communication overhead: small, targeted queries are inefficient due to the communication and parsing involved, while large queries produce large result sets that need to be further processed on the client side.

3.3. Browsing and visualization

The user interface of Flink is a pure Java web application based on the Model-View-Controller (MVC) paradigm. The key idea behind the MVC pattern is a separation of concerns among the components responsible for the data (the model), the application logic (controller) and the web interface (view). The Apache Struts Framework used by Flink helps programmers in writing web applications that respect the MVC

[10] http://www.esri.com/software/arcwebservices/.

pattern by providing abstract application components and logic for the pattern. The role of the programmer is to extend this skeletal application with domain and task specific objects.

The model objects of Flink use the graph model of the JUNG programming toolkit. JUNG[11] is a Java library (API) that provides an object-oriented representation of networks as well as implementations of important measures and algorithms used in (social) network analysis. The model objects loosely map the underlying ontology and retrieve data dynamically from the RDF store as needed for the presentation.[12] The network itself and the most commonly accessed objects are cached to improve performance.

In the view layer, servlets, Java Server Pages (JSP) and the Java Standard Tag Library (JSTL) are used to generate a front-end that hides much of the code from the designer of the front-end. This means that the design of the web interface may be easily changed without affecting the application and vice versa. In the current interface, Java applets are also used on parts of the site to allow the user to interact with the visualization.

We consider the flexibility of the interface important because there many possibilities to present social networks to the user and the best way of presentation may depend on the size of the community as well as other factors. The possibilities range from "text only" profiles (such as in the SNS Orkut[13]) to fully graphical browsing based on network visualization (as in the FOAFnaut[14] browser). The uniqueness of presenting social networks is also the primary reason that we cannot benefit from using Semantic Web portal generators such as HayStack [5], which are primarily targeted for browsing more traditional object collections.

The user interface also provides mechanisms for exporting the data. For more advanced analysis and visualization options, the data can be downloaded in the format used by Pajek, a popular network analysis package[3]. Users can also download profiles for individuals in RDF/XML (FOAF) format. Lastly, we provide marker files for XPlanet, an application that visualizes geographic coordinates and geodesics by mapping them onto surface images of the Earth (see Fig. 1).

4. Social network analysis

The information extraction in Flink is not only the basis of the web application described above, but also provides the data for a sociological study about the role of networks in scientific innovation.

Social network analysis [20,22] is a specialization of the study of networks [1] and it has been applied to a variety of social settings including networks of entrepreneurs, terrorist networks, health (sexual) networks, networks of innovation, etc. Network analysis provides the necessary techniques to prove hypothesis (theories) that link network participation to effects on substantial outcomes such as the performance of an individual or groups of individuals.

A key idea in the structural approach to social science is that the way an actor (an individual or a group) is embedded in a network offers opportunities and imposes constraints on the actor. Occupying a favored position or having preferred kinds of personal connections means that the actor will have better access to valuable information, resources, social support, etc. and will be exceedingly thought after for such opportunities by actors in less favorable positions. In short, social network participation (social capital) might explain a significant proportion of the differences in performance when looking at different, but comparable actors.

With our study of the Semantic Web community our goal is to verify and extend existing theories that relate network participation to innovation in science. In context of the related work (see also Section 5), our methods offer a unique opportunity in terms of the size of the network, the amount of data available and the possibility to observe the dynamics of the network.

A couple of notes are in order about the quality of the data that we obtain, especially in light of using this data for the purposes of social network analysis:

- **Interpretation of the networks**
  One might have noted already that the network obtained from mining the web is a multiplex network on its own, possibly reflecting the co-authorship network, the discussion networks obtained from emails or some other relationship. A closer look at the re-

[11] http://jung.sourceforge.net.
[12] The danger of a close mapping between the ontology and the runtime model is that the application needs to be rewritten whenever the underlying ontology changes.
[13] http://www.orkut.com.
[14] http://www.foafnaut.org/.

sults for a single person (Frank van Harmelen) shows that 44 of the first 100 results returned (from a total of about 10,000) relate to publications and 9 to emails. (Note that the same publication may be referenced in different web pages.) Nevertheless, this network may complement the other networks for different types of relationships (such as informal relationships) and data missing from the other sources (e.g. we may not be aware of all mailing lists related to the Semantic Web).

- **Errors in the extraction of specific cases**

  The network is also bound to contain errors due to the method of collection. The search for co-occurrence is carried out on the syntactic level and shows the typical drawbacks of internet search. For example, it is possible that some of the returned pages are about a different person than the one intended by the query. Ambiguity particularly effects people with common names, e.g. Martin Frank. This danger is mitigated by including the disambiguation term in the query.

  Queries for researchers who commonly use different variations of their name (e.g. Jim Hendler versus James Hendler) or whose names contain international characters (e.g. Jérôme Euzenat) may return only a partial set of all relevant documents known to the search engine.[15] Name ambiguity also effects Google Scholar. For example, the person "York Sure" is identified as a co-author of publications that are published in New York.

  With respect to our use case, the situation is analogous to obtaining incorrect data on a network questionnaire for a part of the respondents, namely those with problematic names. However, this does not represent a problem in computing statistics if the fraction of the cases effected this way remains small.

- **General noise**

  Information extraction will not only effect specific cases, but create a general noise. For example, a co-occurrence of names on a web page need not indicate any social relation in the sociological sense and may be in fact a pure coincidence (e.g. names in a phone directory). Reliability may also be effected by Google itself: the phenomenon of Google Dance can alter the measured association values depending on the time of the query. Such noise in the data, however, will not skew social network statistics as long as it is distributed in an independent manner.

Despite the above difficulties in data collection, we are confident that the quality of the data will allow us to use it for the purposes of network analysis. To verify our method, we also plan to execute a separate study, where we compare the results from a traditional questionnaire method to the acquisition methods described here.

The results of our study of the Semantic Web community may be of interest to both this community and the area of research policy in general, therefore we plan to report on this work in future publications.

5. Related work

Due to the interdisciplinary nature of the work, namely a technological innovation supporting a social science study, the related work is far and wide.

Semantic Web research has produced a number of demonstrations in the area of semantics-based knowledge management, in particular semantic portals for browsing large collections of documents or other objects. Ontology-based knowledge management was the focus of the European On-To-Knowledge project [18] and the more recent SEKT project.[16] The specific area of ontology-based portals has been the subject (among others) of the early work on the SEAL portal generator [16] and the more recent development of the Haystack framework [5]. Flink shares a technological basis and architecture with these projects, with the difference that the "collection" to be presented is a set of persons and the links between them are provided by their social connectivity. The focus on these connections strongly influences the presentation. For example, the ties themselves are presented as individual objects on separate pages. Also, network visualizations (sociograms) are used to orient the user and to provide relevant context information.

In traditional works of scientometrics, scientific networks are investigated by collecting data manually (through interviews or questionnaires), by investigating co-authoring and co-citation in scientific publications

[15] Worthwhile to note that the ambiguity of queries with respect to the content is precisely the problem addressed by Semantic Web technology, in particular FOAF for finding people.

[16] http://www.sekt-project.com.

[6,2] using commercially available databases or by looking for other kinds of evidence of co-participation in research activities, such as public information about project grants [10,8].

Our approach to data collection is part of the more recent trend of applying methods of Computer Science to mining networks from electronic data. As these methods are advanced by computer scientists with an interest in networks, the focus of this literature is clearly on the methods of extraction or analysis rather than the social theory. Emails are the source of social networks in [7,21], while other projects extract networks from web pages with methods similar to ours [14,12] or—somewhat less successfully—by analyzing the linking structure of the web [10]. As first to publish such a study, Paolillo and Wright offer a rough characterization of the FOAF-web in [11].

With our interdisciplinary approach, we hope to contribute both to the methods of network analysis and to the theory of research and innovation. We build our work on the possibilities offered by Semantic Web technology in the collection of data, in particular, the aggregation of information from heterogeneous sources. We complement this with the methodology of social network analysis to learn new insights about the role of the networks in the work of the community, thereby benefitting both network theory and the community under investigation.

Lastly, it is important to note with respect to Flink that the system is applicable to a broader range of communities than the one that is featured in the current application. The few comparative studies in webometrics (web-based scientometrics) suggest that real-world networks of largely academic research communities (such as the Semantic Web community) are closely reflected on the web [10,15]. This suggest that our system could be used to generate presentations of scientific communities in different areas, potentially on much larger scales. With different sources of data, the framework could also be used to visualize communities in areas other than science, e.g. communities of practice in a corporate setting.

6. Conclusions and future work

With the spread of the first computers we believed that as machines replace humans we will interact with them more than with each other, making the world less of a social space. Paradoxically, it seems that nothing could be less true: in the end we shaped our information systems to our form and made them the carriers of efficient forms of communication (from emails to blogs), which allowed us to move much of our social life in the electronic domain.

Our social connectivity might have even increased in importance in the last years simply by the virtue of the information overload we are facing. Browsing the web has become almost futile: the likelihood of finding valuable information by simply following links from page to page has dropped considerably due to the sheer size of the web. Picking up the valuable pieces of information from the mailings lists or blogs that we pretend to follow would require reading them all. That is impossible, unless someone has informed us before about the relevance of an item.

Our social connections not only direct our search in infospace by alerting us to relevant information, but also help to weigh in the authority of the information. When forming a "first impression", the content of a webpage is almost secondary as to how we got there. Was it an email from someone we consider an expert? Was it a link from a website we came to trust? (In fact, this is the thinking behind Google's PageRank algorithm: a webpage is only as authoritative as the ones referring to it.)

If we only had a way to program the underlying reasoning into our machines and provide them with the necessary background information, they could help us much further in distinguishing relevant from irrelevant, trustworthy from corrupt. However, as with most of the content in the electronic domain, almost none of the existing electronic information about our social connections is directly processable to our information systems. Most of it is locked in formats that were not chiefly intended to carry this information. This information needs to be extracted and represented in more formal ways. We may also need these representations to allow the users to enter additional information not directly accessible from an information system. (After all, much of our social life still occurs outside of our systems...)

Thus, the first challenge in the area of social software is the extraction, representation and aggregation of social knowledge. In this article, we have shown how advanced technologies from the Semantic Web domain

(applied information extraction, knowledge representation, ontology mapping) can help in this process. While technology is important, keeping in touch with social science will be just as important in the future. For example, a practical question we encountered in our work concerns the multiplexity of social relations: a relationship between two individuals may have a different significance to different areas of social life. (The most trivial example is the occasional overlap between work and private relations.) Creating a *social ontology* that would allow to classify social relationships along several dimensions is among the future work and so is the finding of patterns for identifying these relationships using electronic data.

In terms of technology, the current bottleneck in scalability is the performance of aggregation (identity reasoning) due to the lack of standard query and rule languages and efficient implementations in RDF stores. Representing context information in a standard and efficient manner will also be necessary to exchange context information among servers.

The extracted and aggregated information, possibly complemented by additional input from the users in the form of a user profile, provides the valuable data needed for adding more intelligence to knowledge intensive applications, in particular improving the navigation of large information stores through collaborative filtering. In our work, the information is constituted by publications and emails, the works and communications of the community. However, networks themselves may also be the focus of interest. Network analysis can benefit communities by identifying the network effects on performance and helping to devise strategies for the individual or for the community accordingly.

In terms of social network analysis, the use of electronic data provides a unique opportunity to observe the dynamics of community development. (This is difficult, if not impossible, with the traditional questionnaire methods of data collection due to the amount of work required from both participants and researchers.) In the future, as our social lives will become even more accurately traceable through ubiquitous, mobile and wearable computers, the opportunities for social science based on electronic data will only become more prominent.

We conclude by noting that the aggregation of social information from disparate sources without permission from the individuals involved is also likely to be the subject of much debate in the future, especially if these sources were originally created for a different purpose, and thus their integration could not have been foreseen. Standard representations, distributed storage and privacy mechanisms should provide the answer by providing protection over one's own social information, but still allowing it to be exchanged with relative ease when required.

Acknowledgement

Funding for this research has been provided by the Vrije Universiteit Research School for Business Information Sciences (VUBIS).

References

[1] A.L. Barabási, Linked: The New Science of Networks, Perseus Publishing, 2002.
[2] A.L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, Evolution of the social network of scientific collaborations, Physica A 311 (3–4) (2002) 590–614.
[3] V. Batagelj, A. Mrvar, Pajek—program for large network analysis, Connections 21 (2) (1998) 47–57.
[4] J. Broekstra, A. Kampman, F. van Harmelen, Sesame: An Architecture for Storing and Querying RDF and RDF Schema, Proceedings of the First International Semantic Web Conference (ISWC 2002), Number 2342 in Lecture Notes in Computer Science (LNCS), Springer–Verlag, 2002, pp. 54–68.
[5] D. Quan, D.R. Karger, How to Make a Semantic Web Browser, Proceedings of the 13th International World Wide Web Conference, New York, USA, 2004, pp. 255–265.
[6] D.J. deSolla Price, Networks of scientific papers: the pattern of bibliographic references indicates the nature of the scientific research front, Science 149 (3683) (1965) 510–515.
[7] P.A. Gloor, R. Laubacher, S.B.C. Dynes, Y. Zhao, Visualization of communication patterns in collaborative innovation networks—analysis of some w3c working groups, CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management, ACM Press, 2003, pp. 56–60.
[8] M. Grobelnik, D. Mladenic, Approaching Analysis of EU IST projects database, Proceedings of the International Conference on Information and Intelligent Systems (IIS-2002), 2002.
[9] P. Haase, J. Broekstra, M. Ehrig, M. Menken, P. Mika, M. Plechawski, P. Pyszlak, B. Schnizler, R. Siebes, S. Staab, C. Tempich, Bibster—a semantics-based bibliographic peer-to-peer system, in: S.A. McIlraith, D. Plexousakis, F. van Harmelen (Eds.), Proceedings of the Third International Semantic Web Conference (ISWC 2004), Hiroshima, Japan, Springer–Verlag, November 2004, pp. 122–136.

[10] G. Heimeriks, M. Hoerlesberger, P. van den Besselaar, Mapping communication and collaboration in heterogeneous research networks, Scientometrics 58 (2) (2003) 391–413.
[11] J.C. Paolillo, E. Wright, The Challenges of FOAF Characterization, Proceedings of the First Workshop on Friend of a Friend, Social Networking and the (Semantic) Web, 2004.
[12] J. Mori, Y. Matsuo, M. Ishizuka, B. Faltings, Keyword Extraction from the Web for FOAF Metadata, Proceedings of the First Workshop on Friend of a Friend, Social Networking and the (Semantic) Web, 2004.
[13] L. Kahney, Making Friendsters in High Places, Wired, July 2003.
[14] H. Kautz, B. Selman, M. Shah, The Hidden Web, AI Magazine 18 (2) (1997) 27–36.
[15] H. Kretschmer, I. Aguillo, Visibility of collaboration on the web, Scientometrics 61 (3) (2004) 405–426.
[16] A. Maedche, S. Staab, R. Studer, Y. Sure, R. Volz, SEAL—tying up information integration and web site management by ontologies, IEEE Data Eng. Bull. 25 (1) (2002) 10–17.
[17] P. Mika, Social networks and the semantic web: the next challenge, IEEE Intell. Syst. 20 (1) (January–February 2005) 82–85.
[18] P. Mika, V. Iosif, Y. Sure, H. Akkermans, Handbook on Ontologies in Information Systems, Chapter 24: Ontology-based Content Management in a Virtual Organization, International Handbooks on Information Systems, Springer–Verlag, 2003, pp. 447–471.
[19] P. Mika, A. Gangemi, Descriptions of Social Relations., Proceedings of the First Workshop on Friend of a Friend, Social Networking and the (Semantic) Web, 2004.
[20] J.P. Scott, Social Network Analysis: A Handbook, 2nd ed., Sage Publications, 2000.
[21] L.A. Adamic, E. Adar, Social Netw. 27 (3) (2005) 187–203.
[22] S. Wasserman, K. Faust, D. Iacobucci, M. Granovetter, Social Network Analysis: Methods and Applications, Cambridge University Press, 1994.