

Application of Semantic Technology for Social Network Analysis in the Sciences

Peter Mika

Department of Computer Science

Vrije Universiteit Amsterdam (VUA)

De Boelelaan 1081, 1081HV Amsterdam, The Netherlands

`pmika@cs.vu.nl`

Tom Elfring

Department of Public Administration and Organization Science

Vrije Universiteit Amsterdam (VUA)

De Boelelaan 1081, 1081HV Amsterdam, The Netherlands

`T.Elfring@fsw.vu.nl`

Peter Groenewegen

Department of Public Administration and Organization Science

Vrije Universiteit Amsterdam (VUA)

De Boelelaan 1081, 1081HV Amsterdam, The Netherlands

`P.Groenewegen@fsw.vu.nl`

July 1, 2005

Abstract

The use of electronic data is steadily gaining ground in the study of the social organization of scientific and research communities, decreasing the researcher's reliance on commercial databases of bibliographic entries, patents grants and other manually constructed records of scientific works. In our work we provide a methodological innovation based on semantic technology for dealing with heterogeneity in electronic data sources. We demonstrate the use of our electronic system for data collection and aggregation through a study of the Semantic Web research community. Using methods of network analysis, we confirm the effect of Structural Holes and provide novel explanations of scientific performance based on cognitive diversity in social networks.

1 Introduction

In social studies of science, research fields have been considered as self-organized communities, also referred to as invisible colleges [13]. Recently smaller groups with shared research interests have been addressed as virtual teams [1, 2]).

Both notions relate to the observation that social connectivity is relevant to many key aspects of research. Conceptualizing research fields as either self-organized communities or virtual teams also allows us to study them using network analysis methods of social science. In past works, various features of the social networks of researchers have proved useful in explaining scientific performance on the individual and organizational level.

The informal nature of scientific communities, however, has also made it difficult to obtain network data for analysis. Traditionally, information about the informal social structure of scientific communities is gathered through labor-intensive methods of data collection, e.g. through interviews or network questionnaires. Alternatively, researchers from deSolla Price [14] to Barabási [4]) have relied on more tangible evidence of formal work relations such as co-authoring and co-citation of scientific pub-

lications or investigated co-participation in research projects [18, 16]. For obtaining much of this data and for measuring scientific performance, most authors have relied on commercially available databases of journal articles, patent- and project grants.

More recently, the use of electronic data extraction is gaining ground in the study of networks. While traditional survey or interview methods are limited in the size of networks and the number of measurements (time-points), electronic data enables large scale, longitudinal studies of networks. In many cases, electronic data also breaks the reliance on commercial providers of information, lowering the costs of access. This process is visible, for example, in the unfolding clash between freely consultable online publication databases based on Information Extraction technology (such as Citeseer¹ and Google Scholar²) and those maintained by the publishers themselves.

Scientific communities in high technology domains such as Artificial Intelligence or Bioinformatics are among the ones that lend most naturally to be studied through their online presence, due the openness of the (academic) research environment and their use of advanced communication technology for knowledge sharing in emails, forums and on the Web. For example, email communication in research and standardization settings are the source of social networks in [15] and [25], while other studies extract social networks from the content of web pages [21, 20] or -somewhat less successfully- by analyzing the linking structure of the Web [18]. As first to publish such a study, Paolillo and Wright offer a rough characterization of the FOAF³ web in [19].

The availability of a multitude of information sources provides the opportunity to obtain a complex view of the social network of a scientific community and thereby it increases the robustness and reliability of research designs. In particular, by decreasing our reliance on the single sources of information our findings become less prone to the

¹<http://citeseer.ist.psu.edu/>

²<http://scholar.google.com>

³FOAF is a popular semantic-based format for describing user profiles and social networks on home-pages. For more information, please see the FOAF project at <http://www.foaf-project.org/>

errors in the individual sources of information. However, the availability of a number of data sources also poses a previously unmet challenge, namely the aggregation of information originating from heterogeneous information sources not primarily created for the purposes of network analysis.

We meet this challenge by offering methodological contributions that also allow us to carry out novel theoretical investigations into network effects on scientific performance.

First, we propose the use of semantic technology in the management of social network data, in particular the semantics-based aggregation of information from heterogeneous information sources. Our system extracts information from web pages, emails and online collections of publications. Semantic technology allows us to uniquely identify references across these sources and merge network data. We will show the validity of this approach by relating network positions to real world status in a scientific community and using our data in proving the well known hypothesis of Ronald Burt with respect to the positive effect of structural holes on innovativeness [10].

Second, we advance the theory of network effects on scientific performance by going beyond a structural analysis of our networks and incorporating the effects of cognitive diversity in ego networks. Our analysis of the potential content of relationships is enabled by our method of extracting the research interests of scientists from electronic sources. In particular, we will look at the content of relations and the effects of cognitive (dis)similarity. We hypothesize that cognitive diversity in the ego network of researchers will be positively related to their performance, especially for newcomers, juniors researchers entering the field.

In the following, we begin by introducing the context of our study, the Semantic Web research community. In Section 3 we introduce our first contribution, our system for extracting and aggregating social network information from various electronic information sources. In Section 4, we formalize our hypothesis concerning the effects of

social and cognitive networks on scientific performance and test these hypothesis using the data collected. We summarize our work in Section 5.

2 Context

The context of our study is the community of researchers working towards the development of the Semantic Web, an extension of the current Web infrastructure with advanced knowledge technologies that have been originally developed in the Artificial Intelligence (AI) community. The idea of the Semantic Web is to enable computers to process and reason with the knowledge available on the World Wide Web. The method of extending the current human-focused Web with machine processible descriptions of web content has been first formulated in 1996 by Tim Berners-Lee, the original inventor of the Web [7].

The Semantic Web has been actively promoted since by the World Wide Web Consortium (also led by Berners-Lee), the organization that is chiefly responsible for setting technical standards on the Web. As a result of this initial impetus and the expected benefits of a more intelligent Web, the Semantic Web has quickly attracted significant interest from funding agencies on both sides of the Atlantic, reshaping much of the AI research agenda in a relatively short period of time⁴

As the Semantic Web is a relatively new, dynamic field of investigation, it is difficult to precisely delineate the boundaries of this network.⁵ For our purposes we have defined the community by including 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 of

⁴Examples of some of the more significant projects in the area include the US DAML program funded by DARPA and a number of large projects funded under the IST initiative of the EU Fifth Framework Programme (1998-2002) and the Strategic Objective 2.4.7 of the EU Sixth Framework Programme (2002-2006).

⁵In fact, it is difficult to determine at what point does a new research concept become a separate field of investigation. With regard to Semantic Web, it is clear that many of the scientists involved have developed ties before their work on the Semantic Web, just as some of the research published in the Semantic Web area has been worked out before in different settings.

2001 (SWWS01), the most significant conference series devoted entirely to the Semantic Web. We note that another commonly encountered way of defining the boundary of a scientific community is to look at the authorship of representative journals (see e.g. [18]). However, the Semantic Web hasn't had a dedicated journal until 2004 and still most Semantic Web related publications appear in AI journals not entirely devoted to the Semantic Web.

The complete list of individuals in this community consists of 608 researchers mostly from academia (79%) and to a lesser degree from industry (21%). Geographically, the community covers much of the United States, Europe, with some activity in Japan and Australia (see Figure 1). As Figure 2 shows, the participation rate at the individual ISWC events have quickly reached the level typical of large, established conferences and remained at that level even for the last year of data (2004), when the conference was organized in Hiroshima, Japan. The number of publications written by the members of the community that contain the keyword "Semantic Web" has been sharply rising since the beginning.

3 Methodology

Our methodology combines existing methods of email and web mining with novel, semantic-based techniques for storing, aggregating and reasoning with social network data. Flink, our self-implemented semantic software supports the complete process of data collection, storage and visualization of social networks based on heterogeneous sources of electronic data. As a Semantic Web application, Flink is the winner of the Semantic Web Challenge of 2004 for the best demonstration of Semantic Web technology in a real application.⁶

The idea of semantic-based representations of user profiles and social networks originates from technology-aware online communities (the blog world), who were

⁶See <http://challenge.semanticweb.org>



Figure 1: Semantic Web researchers and their network visualized according to geography.

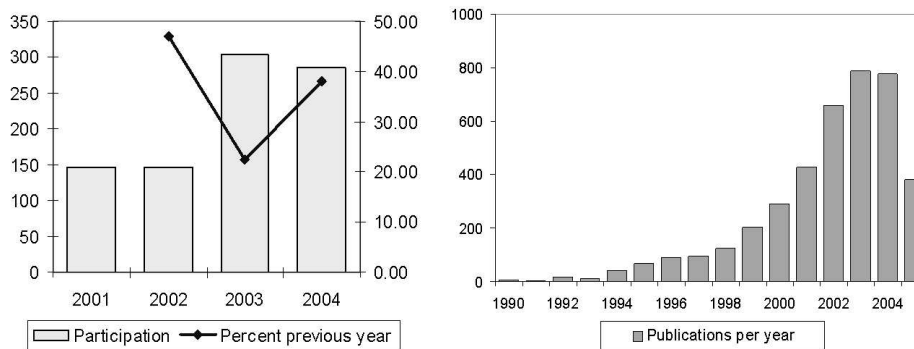


Figure 2: Participation at the international Semantic Web events (2001-2004) and publications per year (1990-2005).

among the earliest adopters of Semantic Web technology. In particular, the format of the Friend-of-a-Friend (FOAF) project was quickly adopted by users, because it allowed to store profiles and social network information linked to the homepages of users, breaking the reliance on profit-oriented social networking services such as Friendster or Orkut. The use of semantic technology also offered the advantage of extensibility: adding new properties to the profiles could be done without breaking compatibility. An example of this is the Speaks-Reads-Writes ontology⁷ for describing what languages the user can read, speak and write.

While semantic technology has been quickly adopted by online communities, it has been left largely unnoticed in the social sciences, despite important benefits for the management of social network data. As Flink demonstrates, semantic technology allows us to map the schema of our information sources and to find correspondences among the instances. The use of standard semantic languages for the representation of social science data makes it possible to use generic Semantic Web tools and infrastructure for editing, storing, querying and reasoning with our data. Lastly, the semantic data store is the basis for a web-based user interface for browsing the data set, computing social network statistics and exporting the networks and the results of the computations.

Figure 3 shows a high level overview of the architecture of the Flink system. The three layers of the system, concerned with data acquisition, representation and visualization, will be introduced separately in the following sections.

We end this Section with a discussion about the use of electronic data for social network analysis. We address both the benefits of electronic data as well as the concerns that can be raised in terms of the reliability of our methods.

⁷<http://www.schemaweb.info/schema/SchemaInfo.aspx?id=48>

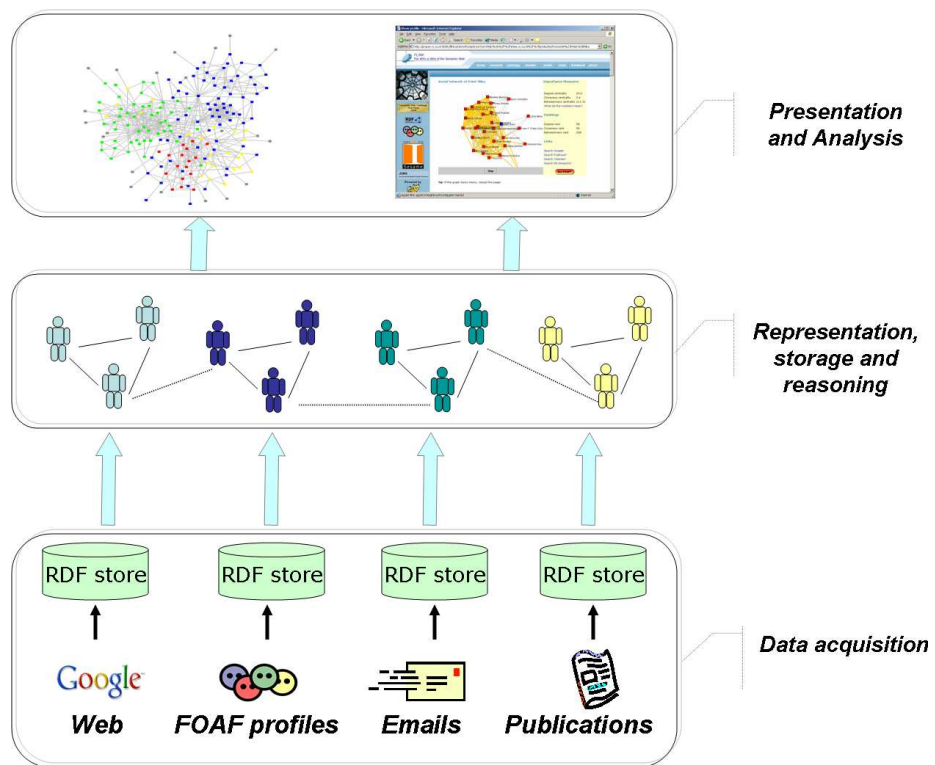


Figure 3: An overview of the architecture of the Flink system.

3.1 Data acquisition

The first layer of the Flink system is concerned with data acquisition. Flink makes use of four different types of knowledge sources: text-based HTML pages from the web, FOAF profiles, public collections of emails and bibliographic data. Information from the different sources is extracted in different ways as described below. In the final step, however, all the data gathered by the system is represented in a semantic format (RDF), which allows us to store heterogeneous data in a single knowledge base and apply reasoning (see the following Section).

The web mining component of Flink extracts social networks from web pages using a co-occurrence analysis technique common in text mining. Although the technique has also been applied before in the AI literature to extract social networks for the automation of referrals (see [21]), to our knowledge this is the first time that the output of the method is subjected to network analysis.

Given a set of names as input, the system uses the search engine Google to obtain the number of co-occurrences for all pairs of names from the membership list of the ISWC community. (The term "(Semantic Web OR ontology)" is added to the queries for disambiguation.) We filter out individuals whose names occurs less than a certain threshold, because in their case the extracted relationships would have very low support.

The absolute strength of association between individuals is then calculated by dividing with the page count of a single individual. In other words, we calculate the fraction of pages where both names are mentioned compared to all pages where an individual is mentioned.⁸ The resulting associations are directed and weighted. We consider such an association as evidence of a tie if it reaches a certain predefined threshold. In our experiments this minimum is set at one standard deviation higher than the mean

⁸We have also experimented with normalization using the Jaccard-formula, but we found that it gives unsatisfactory results if there is a large discrepancy between the web-representation of two individuals. This is the case, for example, when testing the potential relationship between a Ph.D. student and a professor.

of the values, following a 'rule of thumb' in network analysis practice.

The web mining component of Flink also performs the additional task of associating individuals with domain concepts. In our study of the Semantic Web community, the task is to associate scientists with research interests. (The list of terms characterizing research interests has been collected manually from the proceedings of ISWC conferences.) To this end, the system calculates the strength of association between the name of a given person and a certain concept. This strength is determined by taking the number of the pages where the name of an interest and the name of a person co-occur divided by the total number of pages about the person. We assign the expertise to an individual if this value is at least one standard deviation higher than the mean of the values obtained for the same concept.⁹ This is different from the more intricate method of Mutschke and Quan Haase, who first cluster keywords into themes, assign documents to themes and subsequently determine which themes are relevant for a person based on his or her publications [23].

We can map the cognitive structure of the research field by folding the bipartite graph of researchers and research interests.¹⁰ In the resulting simple graph (shown in Figure 7) vertices represent concepts, while an edge is drawn between two concepts if there are at least a minimal number of researchers who are interested in that particular combination of concepts. Note that this is different from the commonly used simple co-word analysis. By using a two-step process of associating researchers to concepts and then relating concepts through researchers we get a more accurate picture of the scientific community. Namely, the names of researchers disambiguate the meaning of words in case a word is understood differently by different authors.

⁹Note that we do not factor in the number of pages related to the concept, since we are only interested in the expertise of the individual relative to himself. By normalizing with the page count of the interest the measure would assign a relatively high score - and an overly large number of interests- to individuals with many pages on the Web. We only have to be careful in that we cannot compare the association strength across interests. However, this is not necessary for our purposes.

¹⁰Bipartite graphs of people and concepts are known as affiliation networks (two-mode networks) in SNA practice. Two-mode networks can be used to generate two simple networks, showing associations between concepts and people [26].

Information from emails is processed in two steps. The first step requires that the emails are downloaded from a mail server and the relevant header information is extracted. In a second step, the individuals found in the collection are matched against the profiles of the members of the target list to filter out relevant profiles from the collection. (See the following Section.)

Although not used in the current experiment, FOAF profiles found on the Web can also be used as an information source. First, an RDF crawler (scutter) is started to collect profiles from the Web. A scutter works similar to an HTML crawler in that it traverses a distributed network by following the links from one document to the next. Our scutter is focused in that it only collects potentially relevant statements, i.e. those containing FOAF information. The scutter also has a mechanism to avoid large FOAF producers that are unlikely to provide relevant data, in particular blog sites¹¹. Once FOAF files are collected, the second step again involves filtering out relevant profiles.

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.¹² An alternative source of bibliographic information (used in previous versions of the system) is the Bibster peer-to-peer network [17], which allows to export bibliographic information in an RDF-based format.

3.2 Representation, storage and reasoning

All information collected through the data acquisition layer is represented in RDF (Resource Description Framework), a lightweight knowledge representation format commonly used in the Semantic Web. RDF is a recommendation from the World Wide

¹¹The overwhelming presence of these large sites also make FOAF characterization difficult. See [19]. We ignore as we don't expect many Semantic Web researchers to maintain blogs and the amount of information would make it difficult to work with the data.

¹²Note 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.

Consortium and was designed specifically for distributed knowledge representation. An RDF document consists of a set of facts expressed as statements in the form of a (subject, predicate, object) triple. For example, to express the fact that the author of a book is a given person, we would insert the triple (book, hasAuthor, person) into our knowledge base. Particular to the web-based purpose of RDF is that the constituents of triples (subjects, predicates and objects) are uniquely identified. This also means that an RDF document can unambiguously refer to entities defined in other documents, wherever they may reside on the Web. This way the World Wide Web can develop into a web of interlinked knowledge bases, which is the vision of the Semantic Web.

Further, RDF is extensible with vocabularies (ontologies) defining specific terms for particular application domains. Ontologies are also expressed in RDF, which means that ontologies can build on terms defined in other ontologies. An example of such a vocabulary is the previously mentioned Friend-of-a-Friend (FOAF) ontology, a simple, generic vocabulary for describing persons and social networks. Another example is the Semantic Web Research Community (SWRC) ontology, which provides the necessary terms for describing publications.

In terms of data management for the social sciences, RDF is a key technology for aggregating information from heterogeneous information sources. The first step in this process is to express all information using a common representation, i.e. RDF. Personal information and social networks are described in FOAF, emails are expressed in a proprietary ontology, while publication metadata is expressed in terms of the SWRC ontology.

After normalizing syntax, the next step is to bridge the semantic gap between the information sources. This consists of mapping the schema and instances of the ontologies used.

Schema matching is a straightforward task in our case. Since the ontologies are known, we can simply insert statements that link corresponding classes in related on-

tologies. For example, we can state that the Author class of the SWRC ontology is a subclass of the Person class of the FOAF ontology, reflecting the rather trivial knowledge that authors of publications are people.¹³

The matching of instances is a more difficult task and it would be close to impossible to automate without the use of advanced knowledge technology.¹⁴ Identity reasoning is required to establish the identity of objects -in our case individuals- across multiple sources of information, based on the fragments of information available in the various sources. An example scenario is shown in Figure 4. Here, the goal is to conclude (based on supporting information) that all three sources describe the same person, even though he is named slightly differently. Note that this is not a single step derivation from the original sources but an iterative process, where the knowledge learned in one step is added to the knowledge base and used in the next round of reasoning. Reasoning steps are repeated until no new facts can be derived. The technical details, however, are beyond the scope of the current publication.

Semantic technology also enables us to reason with social relationships. For example, we have added a rule to our knowledge base which states that the co-authors of publications are persons who know each other. Similarly, the reasoning engine concludes that senders and receivers of emails know each other. In the future, the technology will also allow us to build more refined vocabularies of social relationships, for example to include negative relationships. (The current FOAF ontology only contains a single knows relationship).

3.3 Visualization and Analysis

The web interface of Flink allows visitors to browse and visualize the aggregated information about the social connectivity and professional interests of Semantic Web

¹³That this is may not be the case in the future has been demonstrated recently by a group of MIT students, who have created an application to automatically generate scientific papers. Sadly enough, their works has been accepted at various conferences as legitimate publications.

¹⁴Manual solutions to the problem are completely excluded at the scale of our study.

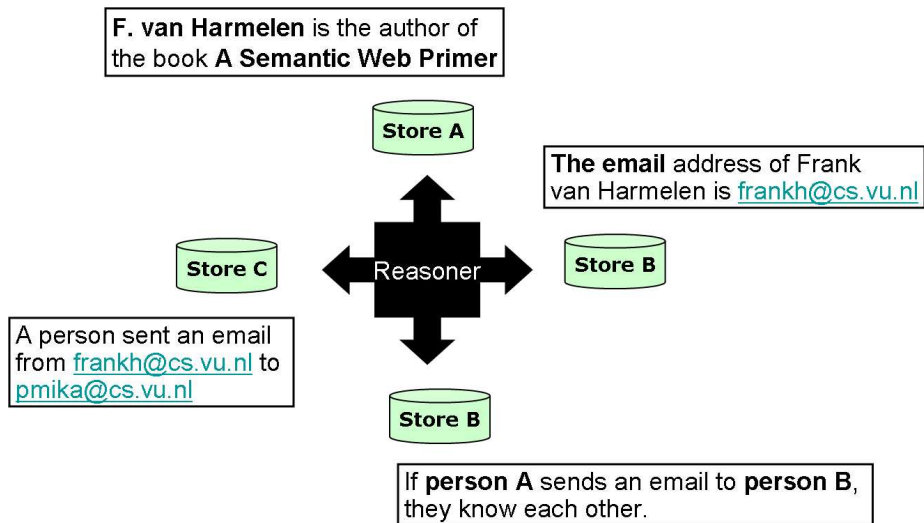


Figure 4: An example of identity reasoning. Among others, the reasoner will conclude in this case that Peter Mika knows the author of the book *A Semantic Web Primer* [3]

researchers. Researchers can also download their profiles in FOAF format. Although it is not possible to edit the information on site, researchers can take the FOAF files provided and store it at their own sites upon editing it. (The new information will be added at the next update of the website when it is found by the FOAF crawler.) The web interface is built using Java technology, in particular the Java Universal Network Graph (JUNG) API. We encourage the reader to visit the website at <http://flink.semanticweb.org>.

Besides visualization, the user interface also provides mechanisms for computing most of the statistics mentioned in this paper. It is also possible to download the network data and statistics for further analysis in the format used by the Pajek network analysis package [5]. Lastly, we provide marker files for XPlanet, an application that visualizes geographic locations and geodesics by mapping them onto surface images of the Earth (see Figure 1).

3.4 Electronic data for network analysis

Electronic data is steadily gaining ground as a basis for carrying out network analysis and it is worthwhile to summarize the general benefits and trade-offs associated with the use of electronic data and the specific characteristics of our method.

We begin by noting that the term electronic data covers information originating from electronic documents or systems of any kind, although most studies focus on internet technology, in particular electronic communication networks and online communities. In the first case, the information sources typically include the structure or content of messages passed through email networks, mailing lists, chats, forums, message boards etc. In the case of online communities, the source of information is chiefly the content or linking structure of documents on the World Wide Web or certain parts of it (e.g. blog communities).

The most important benefits of electronic data acquisition can be summarized as follows:

- **Scalability:** Since data extraction does not require the direct involvement of the subjects of the research, the scale of networks that can be studied increases from tens of subjects (typical for interview methods) to hundreds, thousands or more. This also enables longitudinal designs with an arbitrary number of time points, while traditional studies on network evolution typically rely on data collected at only two distinct time points.
- **Unobtrusiveness, lack of bias:** Electronic data collection is unobtrusive and leaves much less possibility for bias than traditional data collection.
- **Repeatability:** As a matter of principle, the same extraction method applied to the same set of data should always produce the same results. This is difficult, if not impossible to achieve with interview or survey techniques of data collection, but an inherent property of automated information extraction methods. This

means that research results can be reproduced by others at any time.

While these benefits are common to all methods (including ours), it is important to point out the difficulty in contrasting the related work in this area. Different sources dictate different methods for extracting networks from the underlying data and even different extraction methods applied to the same kind of data can lead to significant differences in the results. As a consequence, researchers report conflicting experiences with regard to the success of using electronic data for network analysis.

This has also increased the caution (if not suspicion) of social scientists, rightfully concerned by the quality of data. As an illustration, we would only like to list some of the possible sources of errors in our own method of mining web-based information sources:

- **Multiplexity**

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 results for a single person (Frank van Harmelen) shows that 44 of the first 100 results returned (from a total of about ten thousand) 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 (e.g. project co-participation) 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 vs. 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, which is itself based on Information Extraction technology. A typical error is that a person named "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 will not effect the significance of the results as the fraction of the cases effected this way remains small.

- **General noise**

Errors in information extraction not only effect specific cases, but also 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). The strength of support may also be effected by the coverage and reliability of the search engine. Such noise in the data, however, could only result in a Type II error, i.e. missing significant results in the data.

We believe that these difficulties should not be a source of discouragement as it is possible to validate electronic data and reduce error through triangulation. Validation, as carried out for example in the work of Kretschmer and Aguillo [22], consists of

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

establishing the connection between real world networks in science and their online images. Note that such a validation does not negate the benefits of electronic methods, since the collection of manual data ("the golden standard") only needs to be carried out once and even then on a sample of the population. In our current work, we do not compare our network to any golden standard on a dyadic level, but we will show the correspondence between real world status and centrality in the online network.

We also carry out triangulation by relying on multiple sources of information. Again, this approach is made possible by the co-identification of actors in various sources using semantic technology. Although we cannot ensure that our networks precisely map particular kinds of real world networks, this is not required either. Namely, we are only interested in the effect of the online networks on scientific performance, and not the mechanisms through which networks arise. For our purposes, aggregating network information serves the practical use of increasing the already significant portion of the variance in individual performance that we can explain through online network effects.

4 Results

We have mapped the structure of the Semantic Web community by combining knowledge from three types of information sources as described in Section 3. The actual data set collected on March 17, 2005 contains ties based on 337000 Semantic Web-related web pages¹⁶, a collection of 13323 messages from five mailing lists¹⁷ and 4016 publications from Google Scholar.

The network extracted from e-mail is slightly more similar to publications than

¹⁶This count does not take multiplicity into account, i.e. a webpage may be tied to more than one names. At the time, there were altogether roughly five million pages on the Web where the term "Semantic Web" is mentioned. In general this shows that the community is highly visible on the Web: in comparison, there were about 13 million pages with the term "Artificial Intelligence" and about 1.2 million pages with the term "social networks".

¹⁷These are the rdf-interest, public-swbp-wg, www-webont-wg, public-webont-comments, semantic-web mailing lists, all maintained by the World Wide Web Consortium.

Pearson	e-mail	pub	web
e-mail	1.000		
pub	0.072	1.000	
web	0.064	0.326	1.000

Table 1: Pearson correlations of the three networks extracted from e-mail lists, a publication database (Google Scholar) and web pages.

webpages (especially if we raise the threshold for emails), while webpages and publications are much more correlated. This confirms our intuition that webpages reflect publication activity more than the discussion networks of emails. (Table 1 shows the results of a QAP analysis performed with UCINET.)

In the following, we take the aggregation of the networks as the object of our study, despite the high correlations between the networks. We do so because all networks contain a number of unique ties beyond the overlap. (For example, email lists reveal working group collaborations that may not be manifested in publications.) The aggregated network thus contains a tie between two persons if there is a tie in either the web, email or publication networks. We are not aggregating the weights from these underlying networks as these weights are measured in different units (association weight, number of emails, number of publications), which are difficult to compare.

4.1 Descriptive analysis

Out of the 607 actors, 497 belong to the main component of our network. This connected component itself shows a clear core-periphery structure, supporting our original choice for the boundary definition of the Semantic Web community. (This would not be the case if we would see, for example, two distinct cores emerging.) The single and continuous core/periphery analysis performed with UCINET suggest core sizes of 66 and 114 respectively, where the membership of the larger core subsumes the smaller core with only three exceptions. (The concentration scores show a fairly even decline from their maxima, which suggests that the clusters outside the core are not significant

in size and cohesiveness compared to the core.) The presence of a single, densely connected core also means that the measures of coreness, closeness and betweenness are highly correlated in our network.¹⁸

There is also compelling evidence that measures of the centrality of actors coincide with real-world status in the Semantic Web community. In Figure 5, we have listed the top ranking actors according to our centrality measures and labeled them with their positions held in the community. These positions include chairmanship of the ISWC conference series and editorial board membership at the Journal of Web Semantics¹⁹ and the IEEE Intelligent Systems journal²⁰, the two main sources of Semantic Web-related publications. We also looked at the chairmanship of working groups of the World Wide Web Consortium (W3C), the most influential standards organization in the Semantic Web area.

Ian Horrocks, Dieter Fensel, Frank van Harmelen have been the chairs of the three ISWC conferences held up to date (2002-2004). Stefan Decker and Deborah McGuinness were two of the four chairs of the Semantic Web Working Symposium (SWWS) held in 2001. Rudi Studer and Stefan Decker represent also two of the four editors' in chief of the recently established Journal of Web Semantics. Deborah McGuinness, Frank van Harmelen, Jim Hendler, Jeff Heflin, Ian Horrocks and Guus Schreiber have been co-chairs and/or authors of key documents produced by the Web Ontology (OWL) Working Group of the W3C. Guus Schreiber is also co-chair of Semantic Web Best Practices (SWBP) Working Group, a successor of the OWL group. Jim Hendler is also the current editor-in-chief of the IEEE Intelligent Systems journal. Carole Goble, Toru Ishida and Rudi Studer are joint editors-in-chief of the Journal of Web Semantics.

By looking at the table we can note that all of the common measures of centrality

¹⁸In an ideal C/P structure, betweenness correlates highly with closeness, because actors in the core lie on a large portion of the geodesic path connecting peripheral actors. In other words, peripheral actors have to go through actors in the core to reach each other. Similarly, coreness and closeness correlate because actors in the core are close to each other as well as to actors on the periphery, while peripheral actors are only close to actors in the core, but not to each other.

¹⁹Elsevier, see <http://www.websemanticsjournal.org/>

²⁰IEEE Computer Society, see <http://www.computer.org/intelligent/>

assign high scores to actors with real world status, and we can also ascertain that there are no key position holders of the community whose names would not appear among the first 20 ranks (first three columns). It is also clear that most of the influential members of the community are also successful in terms of the number of publications (fourth column). In terms of impact, i.e. the average number of citations per publication, however, there are members of the community who perform higher than the position holders (fifth column). The explanation is that some peripheral members of the community have highly successful publications in related areas (e.g. agent systems or XML technology). These publications mention the Semantic Web, but are targeted at a different audience than the Semantic Web community.

Indegree		Closeness		Structural Holes		Publications		Impact	
Name	Value	Name	Value	Name	Value	Name	Value	Name	Value
Steffen Staab	119	Ian Horrocks	0.476	Ian Horrocks	113	Steffen Staab	81	Rakesh Agrawal	684
Dieter Fensel	114	Steffen Staab	0.469	Steffen Staab	105	Dieter Fensel	69	Daniela Florescu	191
Stefan Decker	95	Dieter Fensel	0.468	Dieter Fensel	99	Mark Musen	65	David Kinny	180
Enrico Motta	61	Frank van Harmelen	0.467	Frank van Harmelen	91	Ian Horrocks	57	Ora Lassila	166
Frank van Harmelen	59	Stefan Decker	0.458	Stefan Decker	80	Alexander Maedche	53	Honglei Zeng	153
Raphael Volz	59	Rudi Studer	0.438	Rudi Studer	63	Rudi Studer	50	Stuart Nelson	117
Ian Horrocks	55	Enrico Motta	0.434	Guus Schreiber	48	Amit Sheth	47	Michael Wooldridge	91
Sean Bechhofer	48	Sean Bechhofer	0.427	Enrico Motta	44	Katia Sycara	46	Ramanathan Guha	85
Katia Sycara	48	Carole Goble	0.425	Raphael Volz	43	Frank van Harmelen	42	Donald Kossman	83
York Sure	47	Ying Ding	0.424	York Sure	43	Carole Goble	42	Sofia Alexaki	61
Carole Goble	46	Guus Schreiber	0.421	Tim Finin	43	Wolfgang Nejdl	42	Laks Lakshmanan	60
Guus Schreiber	46	York Sure	0.408	Sean Bechhofer	42	Stefan Decker	41	Paolo Atzeni	57
Rudi Studer	46	Peter Crowther	0.407	Katia Sycara	41	Tim Finin	41	Michael Uschold	56
Peter Crowther	40	Alain Leger	0.405	Carole Goble	36	Chen Li	41	Richard Fikes	56
Deborah McGuinness	37	Raphael Volz	0.405	Ora Lassila	27	Enrico Motta	40	Ray Ferguson	55
Ying Ding	35	Herman ter Horst	0.403	Chen Li	26	Nicola Guarino	34	Boris Wolf	53
Jean Francois Baget	34	Jim Hendler	0.401	Richard Benjamins	25	John Domingue	33	Michael Lincoln	50
Jim Hendler	33	David Trastour	0.401	Matthias Klusch	24	Gio Wiederhold	30	Fereidoon Sadri	46
Pat Hayes	32	Richard Benjamins	0.400	Michael Sintek	23	Anupam Joshi	30	Yannis Labrou	45



Figure 5: Centrality in the social network of researchers reflects real world status

Despite the overwhelming presence of the core, we can still observe significant clusters outside the core and there is also some remaining clustering within the core. The analysis of overlapping cliques shows that the largest, most cohesive cluster outside the core is formed by researchers working on semantic-based descriptions of Web Services, in particular members of the DAML-S coalition. The recently popular topic

of Semantic Web Services is rather an application of Semantic Web technology as opposed to the more foundational work on ontology languages (RDF, OWL), which are the main shared interest of those in the core. (The clustering could be partly also explained that many of the senior researchers have a background in agent-based systems and have worked together in the past in that area.) To show that this is clearly a topic-based cluster, we have mapped the association of researchers with the concept "DAML-S" against the social network. As Figure 6 clearly illustrates, most of these researchers belong to a relatively densely connected subgroup outside the core. (For more information on the method we use to associate researchers with research ideas, please refer to Section 3).

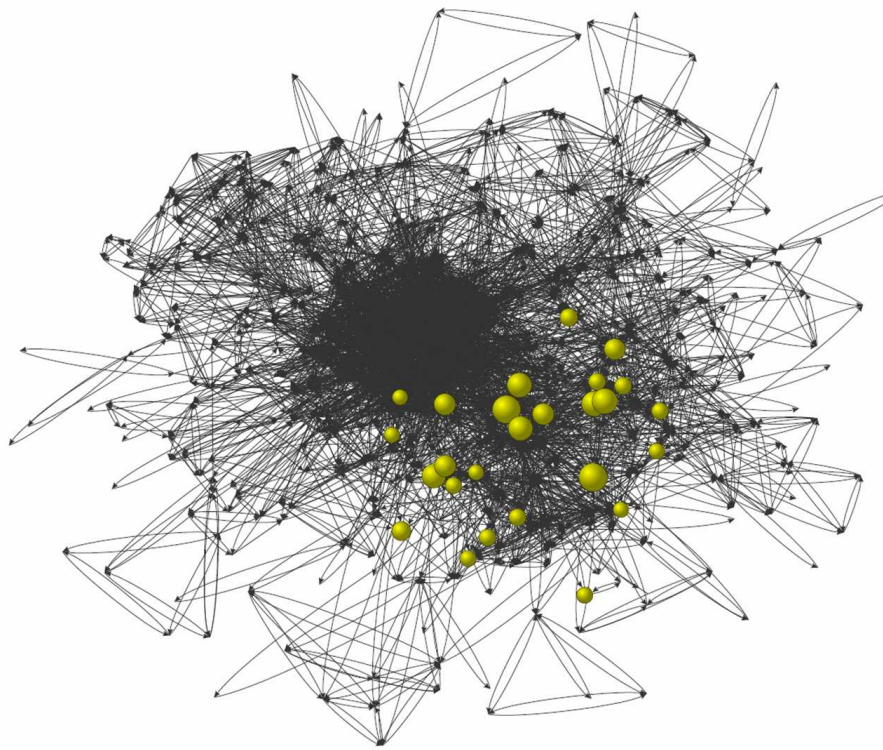


Figure 6: Researchers associated with the concept DAML-S form a cluster outside of the core.

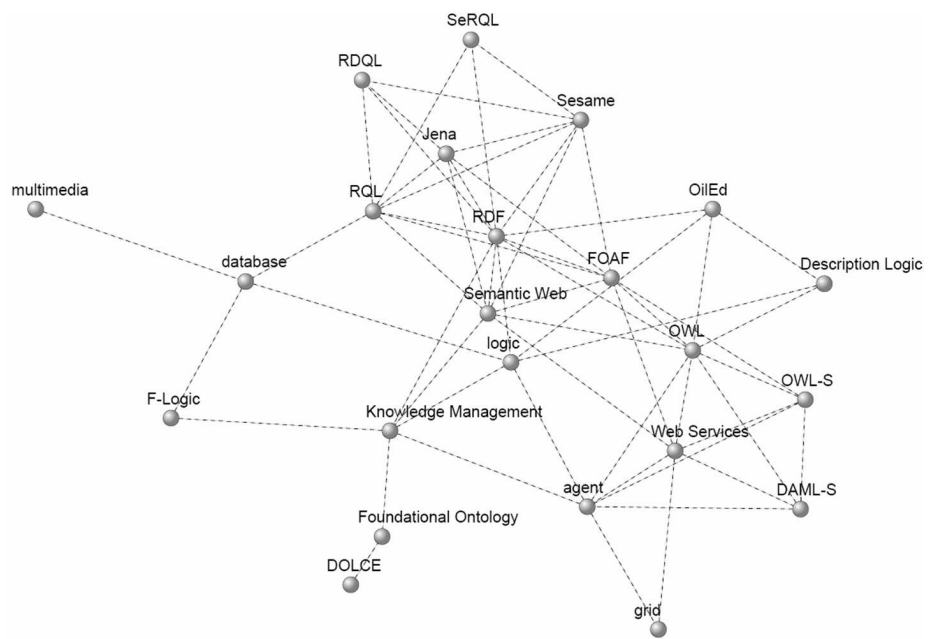


Figure 7: The cognitive structure (ontology) of research topics.

4.2 Structural and cognitive effects on scientific performance

The social network literature has debated the effect of structure on performance from the perspective of the effects of close interconnectedness versus a sparser network [11]. The basic arguments for the positive effects of a dense interconnected network are that these ties foster trust, identification and these combined lead to an easier exchange of information [12]. Opposite this argument stands the argument of diversity, by incorporating ties with diverse other groups through the occupation of a structural hole position more new ideas may be encountered and incorporated into ones work [9].

In small scale situations it has been shown that communication ties that bridge a variety of different groups lead to higher performance as did network density [27]. In a study of researchers working on the development of software, centrality measures have been shown to correlate with scientific productivity. An analysis of email messages in this group (about 50 members) showed that centrality correlated strongly with scientific performance [2]. Centrality was found to be partly, but not completely, a consequence of functional characteristics of the researchers in the field. The reason centrality influences performance is suggested to be a consequence of the benefits a specific individual has from being receiver of a larger amount of (diverse) information.

In the following, we formulate our hypothesis concerning the effects of social networks on scientific performance. First, we test for the effect of ego-network structure, namely the size and density of ego networks. (As previously mentioned, the size of the ego network is also highly correlated with the centrality of the individual.) Second, we look for the additional effects of cognitive diversity in the ego network.

Our primary measure of performance is the number of publications by a given researcher that has received at least a minimum number of citations (dependent variable TOPPUBLI). Based on the general distribution of the number of citations of all publications, we set this minimum at four, excluding 40% of all publications. (The exclusion of minimally cited publications is a commonly used method for filtering out

publications that are unlikely to contain innovative results.) Our second, alternative independent variable is the average number of citations per publication (IMPACT). This is a different dimension than the number of publications, because the same impact can be achieved with less or more publications.

We chose the term performance instead of innovativeness when discussing our dependent variables, because publication-based measures offer a very limited way of measuring genuine innovativeness. For example, an article with innovative content may attract few citations if it is published in a journal with limited visibility. On the other hand a survey or review article with minimal innovativeness may attract a large number of citations due to its very nature, especially if published in a high-visibility journal. (These same arguments can also be raised against the usefulness of the impact factor for measuring journal performance.) Nevertheless, publication and citation numbers are still often used in real life to evaluate the performance of researchers.

In reporting our results, we control for experience as we are interested in the unique role of networks in explaining scientific performance. Experience is a personal factor (external to networks) that naturally correlates with both our network variables and our measures of performance.

We measure experience in years of research within the Semantic Web domain, based on the first publication in the area. Therefore, our measure of experience does not include (possibly extensive) research experience in other domains, before or in parallel to involvement in Semantic Web research. Further, we do not consider the case of researchers who give up Semantic Web research (or researcher altogether). However, we expect this to be quite rare at this point in time.

4.2.1 The effect of network structure on performance

A closer examination of Ronald Burt's measure of effective size reveals that a structural hole has two distinct components: the size and efficiency of the ego-network [8].

Experience	All	<= 5	<= 4	<= 3	<= 2
TOPPUBLI	0.665	0.494	0.350	0.299	0.376
sign.	0.000	0.000	0.000	0.000	0.000
df	496	318	248	172	92
IMPACT	0.009	0.286	0.195	0.254	0.269
sign.	0.848	0.000	0.002	0.001	0.009
df	496	318	248	172	92

Table 2: DEGREE controlled for NEWEXP

In fact, in the most simple case of a network with undirected, unvalued ties, Burt's measure results to be equal to the number of direct ties of an actor (degree) minus the clustering coefficient scaled by a factor of $n - 1$, where n is the number of nodes.

In the following, we examine the separate contribution of these components.

Hypothesis 1a. The number of ties to alters is positively related to performance.

Figure 2 shows the partial correlation between the number of ties of an individual and our dependent variables, controlling for experience. The first column of the table shows the results for all cases, while the columns to the right show correlations for sections of the populations with less than five, four, three or two years of experience.²¹

In general, we can note that the number of ties explains a significant portion of the publication performance measured in the number of publications. A large social network is thus either the cause or the effect of publishing activity.

In the general population, however, degree is not significantly related to impact. In other words, high impact does not necessarily require a large social network and vice versa. However, it also seems that younger researchers are still able to turn the informational advantages of social access into a higher impact of their publications.

Hypothesis 1b. A dense network of ties among alters (closed network) is negatively related to performance.

As expected, clustering in the ego-network of the individual is negatively related

²¹If we look at the separate effects of in- and out-degree instead of the full size of the ego network, we find that in-degree is more correlated with the number of publications, while out-degree is more correlated with impact. (Due to limitations of space, we omit the data.) Unreciprocated ties resulting from the analysis of web pages represent a relationship that is more important for the sending actor than the receiving actor.

Experience	All	<= 5	<= 4	<= 3	<= 2
TOPPUBLI	-0.146	-0.129	-0.200	-0.179	-0.239
sign.	0.001	0.017	0.001	0.013	0.015
df	525	342	267	190	101
IMPACT	-0.066	-0.080	-0.072	-0.144	-0.205
sign.	0.128	0.140	0.236	0.047	0.037
df	525	342	267	190	101

Table 3: CLUSTER controlled for NEWEXP

to publication performance when measured in the number of publications. A dense network is thus an inefficient network as far as publishing is concerned.

The evidence for the negative effect of clustering on the impact of publications is much weaker. It seems that while clustering negatively impacts the number of publications, it has a much smaller effect on impact. We postulate that publications created in a dense network can still have a relatively high impact within a sub-community of researchers.

4.2.2 The effect of cognitive network structure on performance

Burt's measure of structural holes ignores the actual content that moves through the connection provided. More precisely, Burt assumes that different, unconnected sub-groups provide unique knowledge to the broker between them, leading to an advantageous position also in terms of access to knowledge. However, there are more direct ways to establish the link between accessing a diversity of knowledge sources and the performance of the individual.

In a number of previous studies the structural hole argument has been translated to the basic idea of a range of informational sources [24]. The manner in which this variety has been constructed differs in the studies that appeared until now. One example is a study in which organizational variety is taken as a proxy for diversity. Baum et al. considered companies, government agencies and firms with different industrial backgrounds as providing variety [6]. In their study on knowledge transfer Reagans

and McEvily used a functional description of roles and variety of expertise [24].

In a relatively homogeneous research field, we expect that cognitive differences may drive the process of innovation. We expect that differences in the research profile of the ego and his alters may benefit the individual in addition to the already proven positive effect of a large and efficient social network.²²

Hypothesis 2a. Access to cognitive diversity through networks is positively related to performance, especially for younger researchers.

In the following, we measure the cognitive diversity in the ego-network by looking at the difference between the research interests of the ego and his or her alters. We will say that a (structural) tie is a content tie, if there is at least one interest of the alter that is not a current research interest of the ego. We measure diversity by counting the number of content ties of an ego (content-degree). We stipulate positive effects on scientific performance in particular for younger researchers. We believe that senior researchers would be less susceptible to content effects as they can rely on junior researchers in their network (positional advantages) and their functional ties for greater publication performance.

To show the unique contribution of cognitive diversity towards explaining scientific performance, we perform a linear regression with experience, degree and content-degree as predictors (independent variables) and the number of publications (TOP-PUBLI) and average citation (IMPACT) as the outcome (dependent) variables. (We limit our investigations to scientists with at most four years of experience.) The results, shown in Figure 8, indicate that the unique contribution of content degree is significant in both cases. (In fact, in the final models the coefficient of degree is not significantly different from zero any more.) We also find evidence that access to cognitive diversity

²²Note that while we take the cognitive structure as a given, related work by Mutschke and Quan Haase looks at social network-based explanations for the development of the cognitive structure of scientific communities [23]. The authors suggest that the most connected actors (actors with a higher degree centrality) are likely to work on the more central research themes. Renner hypothesizes that the opposite is also true, namely that new ideas are likely to originate from the most peripheral actors. However, such a hypothesis is difficult to prove or refute in practice: the boundary of a scientific network is always fuzzy and a peripheral actor may have many connections to actors outside of the community under investigation.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
(Constant)	0.736‡ (0.000)	0.883‡ (0.000)	-0.635† (0.041)	2.452‡ (0.000)	2.946‡ (0.000)	1.106 (0.265)
DEGREE	0.057‡ (0.000)	-0.036 (0.190)	-0.045 (0.073)	0.144‡ (0.000)	-0.239‡ (0.003)	-0.250‡ (0.002)
CONTENTD		0.292‡ (0.000)	0.293‡ (0.000)		1.210‡ (0.000)	1.211‡ (0.000)
NEWEXP			0.737‡ (0.000)			0.893† (0.027)
R^2	0.106	0.176	0.316	0.074	0.205	0.228

Figure 8: Results of linear regression with the number of publications (TOPPUBLI) as dependent variable (Model 1-3) and the average number of citations (IMPACT) as dependent variable (Model 4-6). N=172, † $p < 0.05$, ‡ $p < 0.01$

has a particularly large effect on the impact of the publications.

5 Conclusions and Future Work

With our interdisciplinary approach to the study of scientific communities, we are aiming to contribute to both the methods of network analysis and the social theory of research and innovation.

In our methodology, we build on the possibilities offered by Semantic Web technology in the aggregation of the data that we have collected from a number of freely accessible online information sources. The use of freely available electronic data (web pages, publications, mailing lists) not only lowers the cost of studying science communities, but also enables us to significantly increase the scale and longitude of our studies. Further, the reuse of multiple information sources allows us to gain a more complete picture of the community under investigation. Semantic technology is crucial for dealing with the arising heterogeneity.

With respect to our method, we note that it is applicable to a broader range of communities than the one featured in the current study. The few existing comparative studies in webometrics (web-based scientometrics) suggest that real-world networks of

largely academic research communities are closely reflected on the Web [18, 22]. This suggest that our system could be used to generate networks 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. 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.

In the above, we have shown the immediate benefits of our methodology by applying it toward a network study of the Semantic Web community. Based on our data set, we have proved the positive effects of a large, efficient (sparse) network on the innovativeness of researchers, confirming the benefits attributed to Structural Holes [9]. We have extended the well-known structural analysis of this scientific community with a novel analysis of the content of relationships. We have shown that diverse cognitive networks have a positive impact on performance beyond the structural effects. Our measure of content degree results to be a much better predictor than simple degree for both the number of publications and the average number of citations.

We are planning to extend our work in this direction, e.g. by investigating whether cognitive diversity in the ego network could have a negative effect in cases where the distance between research areas is overly large. We are also developing measures to study the expected positive effect of achieving a diverse cognitive network with a minimal investment in social ties. Planned improvements to our information retrieval methods should also enable us in the future to determine more precisely the interests of individual researchers in the community. Lastly, we are tracking the development of the Semantic Web community over time using our electronic methods of data collection, providing a wealth of data for future work.

6 Acknowledgement

Funding for this research has been provided by the Vrije Universiteit Research School for Business Information Sciences (VUBIS). The authors would also like to thank Maurits de Klepper for his valuable comments to this paper.

References

- [1] Manju K. Ahuja and Kathleen M. Carley. Network structure in virtual organizations. *Organization Science*, 10(6):741–757, 1999.
- [2] Manju K. Ahuja, Dennis F. Galletta, and Kathleen M. Carley. Individual Centrality and Performance in Virtual R&D Groups: An Empirical Study. *Management Science*, 49(1):21–38, 2003.
- [3] Grigoris Antoniou and Frank van Harmelen. *A Semantic Web Primer*. MIT Press, 2004.
- [4] A.L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311:590–614, 2002.
- [5] Vladimir Batagelj and Andrej Mrvar. Pajek - Program for Large Network Analysis. *Connections*, 21(2):47–57, 1998.
- [6] Joel A. C. Baum, Tony Calabrese, and Brian S. Silverman. Don't go it alone: alliance network composition and startups' performance in Canadian biotechnology. *Strategic Management Journal*, 21:267–294, 2000.
- [7] Tim Berners-Lee, Mark Fischetti, and Michael L. Dertouzos. *Weaving the Web : The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. Harper San Francisco, 1999.

- [8] Stephen P. Borgatti. Structural Holes: Unpacking Burt's Redundancy Measures. *Connections*, 20(1):35–38, 1997.
- [9] Ronald Burt. Structural Holes and Good Ideas. *American Journal of Sociology*, 110(2):349–400, 2004.
- [10] Ronald S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1995.
- [11] Ronald S. Burt. The network structure of social capital. *Research in Organizational Behaviour*, 22:345–423, 2000.
- [12] James S. Coleman. Social Capital in the Creation of Human Capital. *American Journal of Sociology*, 94:95–120, 1988.
- [13] Diana Crane. Transnational networks in basic science. *International Organization*, 25:585–601, 1971.
- [14] Derek J. deSolla Price. Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science*, 149(3683):510–515, 1965.
- [15] Peter A. Gloor, Rob Laubacher, Scott B. C. Dynes, and Yan Zhao. Visualization of communication patterns in collaborative innovation networks - analysis of some w3c working groups. In *CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 56–60. ACM Press, 2003.
- [16] Marko Grobelnik and Dunja Mladenic. Approaching Analysis of EU IST Projects Database. In *Proceedings of the International Conference on Information and Intelligent Systems (IIS-2002)*, 2002.
- [17] Peter Haase, Jeen Broekstra, Marc Ehrig, Maarten Menken, Peter Mika, Michal Plechawski, Pawel Pyszlak, Björn Schnizler, Ronny Siebes, Steffen Staab, and

- Christoph Tempich. Bibster — a semantics-based bibliographic peer-to-peer system. In Sheila A. McIlraith, Dimitris Plexousakis, and Frank van Harmelen, editors, *Proceedings of the Third International Semantic Web Conference (ISWC 2004)*, pages 122–136, Hiroshima, Japan, November 2004. Springer-Verlag.
- [18] Gaston Heimeriks, Marianne Hoerlesberger, and Peter van den Besselaar. Mapping communication and collaboration in heterogeneous research networks. *Scientometrics*, 58(2):391–413, 2003.
- [19] John C. Paolillo and Elijah Wright. The Challenges of FOAF Characterization. In *Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and the (Semantic) Web*, 2004.
- [20] Junichiro Mori and Yutaka Matsuo and Mitsuru Ishizuka and Boi Faltings. Keyword Extraction from the Web for FOAF Metadata. In *Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and the (Semantic) Web*, 2004.
- [21] Henry Kautz, Bart Selman, and Mehul Shah. The Hidden Web. *AI Magazine*, 18(2):27–36, 1997.
- [22] Hildrun Kretschmer and Isidro Aguillo. Visibility of collaboration on the web. *Scientometrics*, 61(3):405–426, 2004.
- [23] Peter Mutschke and Anabel Quan Haase. Collaboration and cognitive structures in social science research fields. *Scientometrics*, 52(3), 2001.
- [24] Ray Reagans and Bill McEvily. Network Structure and Knowledge Transfer: The Effects of Cohesion and Range. *Administrative Science Quarterly*, 48(2):240–267, 2003.
- [25] How to search a social network. Lada Adamic and Eytan Adar. Submitted to *Social Networks*, 2004.

- [26] Stanley Wasserman, Katherine Faust, Dawn Iacobucci, and Mark Granovetter. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
- [27] Ezra W. Zuckerman and Ray E. Reagans. Networks, Diversity, and Performance: The Social Capital of Corporate R&D Teams. *Organization Science*, 12(4):502–517, 2001.