Social Networks and the Semantic Web

Peter Mika
Department of Business Informatics and
Department of Management & Organizations
Free University Amsterdam, The Netherlands
pmika@cs.vu.nl

Abstract

A formal, web-based representation of social networks is both a necessity in terms of infrastructure as well as a prominent application for the Semantic Web. In this paper we present three advances in exploiting the opportunity of semantically-enriched network data: (1) an ontology for the representation of social networks and relationships (2) a hybrid system for online data acquisition that combines traditional web mining techniques with the collection of Semantic Web data (2) a case study highlighting some of the possible analysis of this data using methods from Social Network Analysis, the branch of sociology concerned with relational data.

1. Introduction

The past year has seen the emergence of social networking sites as some of the most popular places on the web, with the first-mover Friendster[1] attracting over 5 million registered users in the span of a few month. These sites allow users to post a profile with basic information, to invite their friends and to link to their profiles in the system. The system also lets the users 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.

Despite the early popularity, users have soon discovered a number of drawbacks to such centralized systems. First, the profiles stored in these systems cannot be exported in machine processable formats, which means that the information is not portable among social networking sites. This became a problem after a number of Friendster alternatives appeared and the users had to recreate their profile -and keep it updated- separately at the different sites. Second, centralized sites did not allow users to control the information they provide on their own terms. Although follow-ups offer several levels of control (e.g. public information vs. only for friends), users often still find out the hard way that their information was used in ways they did not foresee.

Both of these problems can be addressed with the use of Semantic Web technology. The Friend-Of-A-Friend (FOAF) ontology is a first attempt at a formal, machine processable representation of user profiles and friendship networks [4]. Unlike with Friendster and similar sites, FOAF profiles are created by the individual user and stored in a distributed fashion (typically posted on the personal web page of the user). Much like web pages, these profiles also link to the profiles of friends, creating the so-called FOAF-web. In effect, the FOAF-web is a single social network described in a universal format that is directly accessible to machines.

The popularity of social networking on the web and the explosive combination with web semantics open up vast and so far unexplored opportunities for social intelligence on the Web.[2] When social network data is available in machine-processable format such as RDF, network data of online communities does not need to be collected any more by potentially error-prone methods of screen-scraping. Further, it becomes relatively easy to integrate this data on a semantical level with other sources of information. Regardless whether one approaches this data with a commercial or research interest, this means higher quality information that is more amenable to analysis.

This paper presents two advances towards exploiting this opportunity. In Section 2, we introduce a novel system for collecting social network data, which acquires data both from the traditional web and the Semantic Web. Secondly, we exploit this system for a study of the Semantic Web research community using the traditional methods of So-

1 `http://www.friendster.com`

2 FOAF is already the most popular application of Semantic Web technology. Experiments in the UK AKT projects have revealed that over 90 percent of all RDF data on the Web is FOAF. Source: presentation by Nigel Shadbolt, director of AKT.

Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)
0-7695-2100-2/04 $ 20.00 IEEE

cial Network Analysis (SNA) [12, 13] as described in Section 3. Lastly, we summarize the results of this experiment and place it in the context of related and future work in Section 4.

2. Architecture and experiment design

In this section we briefly present the system developed for our experiment (Section 2.1). We also discuss the setting of our case study involving the Semantic Web research community and reflect on some of the concerns and advantages of using our a system for the purposes of network analysis (Section 2.2).

2.1. System design

The system developed for collecting and analyzing online social networks is one of very few hybrid applications that combine methods of mining the traditional, human-oriented web with exploiting the Semantic Web. (See Figure 1 for an overview of the architecture of the system.)

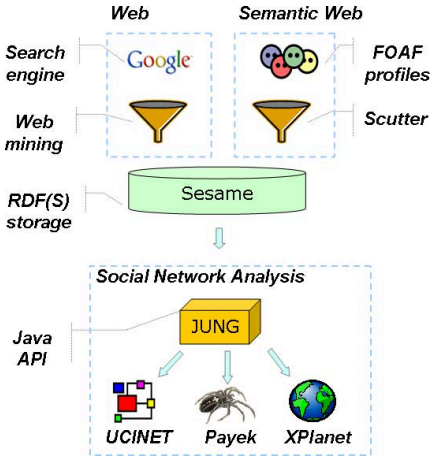

Figure 1. An architectural overview of the system.

Given a set of names, the system uses the index of Google[3] to generate social network data. This method works by submitting all possible pairs of names to the search engine and counting the number of pages returned, i.e. the pages where both names occur. Tie strength is then calculated by normalizing this result with the number of pages returned for the names individually. (This is also known as the Jaccard coefficient [11].)

3 www.google.com

Given a set of names and a list of research topics[4], a related component performs a calculation of Google Mindshare[5] to determine whether a given person is associated with a given research area or not. The strength of this association is normalized only with the pagecount of the person involved. (It is expected that more popular topics attract more people.) The threshold is chosen as the mean plus the standard deviation of the values.

Both the network ties and the topic associations are represented in RDF using a compatible extension of the FOAF model [4]. These extensions are necessary to record the provenance and timestamp of the data collected as well as the weight of the relationships, when available. The RDF data is stored in Sesame, an RDF storage and query facility[6]. Note that since the model is a compatible extension of FOAF, the 'ontologized' network data can be read as a set of FOAF profiles and processed by any FOAF-compatible tool.

Using the Semantic Web, social network data in the form of FOAF profiles is collected by crawling. The system uses a so-called scutter, a robot that traverses the RDF-web by following rdfs:seeAlso links. After recording the provenance and attaching a time-stamp, the knowledge found is added to the Sesame store as well. (At present no automated smushing[7] is performed.)

In a next step, the data in the Sesame server is further enriched by adding the geographical locations of place names found in the FOAF profiles. This step is also carried out automatically by interacting with the Place Finder Sample Web Service[8] that provides geo-location services for over three-million place names.

The data is further processed by the JUNG programming toolkit for (social) network analysis. JUNG is a Java library that provides an object-oriented representation of networks. JUNG also contains the algorithms for calculating several basic network measures and makes it possible to build complex applications using the provided visualizations. (We will show an example of such an application in Section 3.3).

Advanced visualizations are crucial for theory-building in network analysis. For this reason and for performing more sophisticated analysis, the network data was exported from JUNG in the formats used by the Pajek[9] and UCINET[10] packages for Social Network Analysis. (Both tools offer further export to formats used by 3D vi-

4 Or any other kind of concept
5 http://hacks.oreilly.com/pub/h/199
6 http://www.openrdf.org
7 Smushing refers to the aggregation of FOAF data based on uniquely identifying properties, such as email address.
8 http://arcweb.esri.com/arcwebonline
9 http://vlado.fmf.uni-lj.si/pub/networks/pajek/
10 http://www.analytictech.com/ucinet.htm

sualizations.) The JUNG representation was also the basis for generating geographic visualizations using XPlanet[11]. The analysis of these images has been omitted from the present paper for reasons of space.

2.2. Experiment design

With the system introduced above, we collected data on the network of Semantic Web researchers during March, 2004. For our purposes, we defined the network boundary as including only those researchers whose names have appeared at least once as either program committee members or organizers of any of the three international Semantic Web events organized so far ($N = 167$).[12] This means that our focus is only on high-profile researchers, who are likely to be the most committed to this area of research. Unfortunately only a small fragment of this community has a published FOAF profile, which means that we had to rely on traditional web data for the current experiment.

Here, we would like to make a number of observations before treating web data as an input for social network analysis.

First, using the traditional questionnaire and interview methods of SNA, the notion of tie strength is treated as an aggregate measure, reflecting the multi-dimensionality of this concept. In fact, the simple co-occurrence relationship of this network may reflect different types of relationships, such as co-authorship, co-participation at events etc. 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.[13] Although this hypothesis has not been tested, this suggest that the network may show a correlation to the co-authorship network of the researchers, taking multiplicity into account. (Popular publications are mentioned more often.)

Secondly, the data is bound to contain errors due to the method of collection. The search for co-occurrence is carried out on the syntactic level and show 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). Furthermore, searches for researchers who use different variations of their name (e.g. Jim Hendler vs. James Hendler) or international characters (e.g. Jérôme Euzenat) may return only a partial set of all relevant documents known to the search engine. Note that the ambiguity of web searches with respect to the content is precisely the problem addressed by Semantic Web technology.

Lastly, concerns can be raised about the sensitivity of our analysis to the definition of the network boundary, i.e. the initial choice of researchers. While this is a well-known problem in SNA, we have experimented with the sensitivity of some of the measures w.r.t. the size of the network. When removing nodes from the network (and thus simulating the effect of missing information), we have found, for example, that the value of closeness centrality (see following section) remained highly correlated to its original value even when removing a significant number of nodes. For example, on a network of 39 nodes correlation still reached 0.8 after removing a third of all nodes. This gives us sufficient confidence in the reliability of our results.

3. Analysis

In this section we present some of the analysis we have carried out in terms of global network measures (Section 3.1, ego-network measures (Section 3.2) and subgroup analysis (Section 3.3). All the techniques used in this Section are part of the standard toolkit of Social Network Analysis and described in a number of textbooks (e.g. [12, 13]).

3.1. Global network measures

As discussed before, the system we have developed provided a continuous value for the tie strengths between the researchers. By looking at this distribution, we have found that it follows a power-law ($R^2 = 0.9848$), with a mean of 10 and a standard deviation of 15.3. We have decided to analyze two networks by choosing two different thresholds for tie strength.

After deleting the ties below the first threshold of 30 and removing social isolates, the network contained 124 nodes and six components, with a large component of size 114 and five pairs of individuals. In the following, we will refer to this component as the entire network.

Following the same procedure with a higher threshold of 50, the network was left with 93 nodes and 15 components, with a large component of 37. As this component can be considered as the strong core of the entire network, we will refer to it in the following as the core of the network.

The weak network had an average separation of 4 degrees, with the degree distribution following a power law ($R^2 = 0.5673$). This finding is in accordance with the Barabási model of small worlds [10, 2], which predicts the average separation to be on the order of $log(N)$.

When laid out graphically, the image of the network suggests an underlying core-periphery structure to the network (see Figure 2). Nodes with the highest centrality make

11 `http://xplanet.sourceforge.net/`

12 The Semantic Web Working Symposium (SWWS) of 2001 and the First and Second International Semantic Web Conferences (ISWC) held in 2002 and 2003.

13 The rest is divided between pages of events (23), email discussions (17) and miscellaneous items (homepage, teaching, CV, news, listings, directories, search results).

up a single connected cluster in the graph. This combined with our subgroup analysis implies that there is a dense core (with possible several concentric circles of influence) surrounded by a number of smaller sub-communities connected more loosely to this core (see also Section 3.3). We hope that after we will be able to observe this network over time we can determine conclusively the existence of a Semantic Web elite and the options for social mobility.

3.2. Ego-centered network measures

A key idea in the structural approach to social science is that the way an actor is embedded in a network offers opportunities and imposes constraints on the actor. Occupying a favored position means that the actor will have better access to information, resources, social support etc. and will be exceedingly thought after for such opportunities by actors in less favorable positions.

In particular, power and influence in informal networks stem from occupying positions that are central to the network. The measure of centrality has been the subject of several studies in SNA and a number of operationalizations have been proposed.

A simple, but effective measure is the *degree centrality* of the node [12, 13]. Degree centrality equals the graph theoretic measure of degree, i.e. the number of (incoming, outgoing or all) links of a node. This measure is based on the idea that an actor with a large number of links has wider access to the network, less reliant on single partners and because of his many ties often participates in deals as a third-party or broker. Degree centrality does not take into account the wider context of the ego and nodes with a high degree may in fact be disconnected from large parts of the network. However, the degree measure features prominently in the scale-free model, which makes it an important measure to investigate (see above).

A second, more intuitive measure of centrality is *closeness centrality*, which is obtained by calculating the average (geodesic) distance of a node to all other nodes in the network. Closeness centrality thus characterizes the reach of the ego to all nodes of the network.

Two other measures of power and influence are related to the advantage gained through brokering. *Betweenness centrality* measures the extent to which other parties have to go through a given actor to conduct their dealings. Consequently, betweenness is defined as the proportion of paths -among the geodesics between all pairs of nodes- that pass through a given actor. Figure 3 shows the ten highest ranking nodes of the network in terms of degree, closeness and betweenness centrality. For comparison, we also included the highest ranking nodes by the measure of page count (the number of pages returned by Google for an individual) for both the entire network of 167 nodes and the strong core of 37 nodes.

The more complex measure of Ronald Burt is related to the idea of Structural Holes [6]. A Structural Hole occurs in the space that exists between closely clustered communities. According to Burt, a broker gains advantage by bridging such holes. (Brokers are even claimed to have better ideas due to their extended vision [5].) Therefore this measure favors those nodes that connect a significant number of powerful, sparse linked actors with only minimal investment in tie strength.[14] Figure 4 shows the ten highest ranking nodes of the network according to Burt's measure.

Rank	Name	Value
1	Stefan Decker	0.251
2	Andreas Hotho	0.273
3	Rudi Studer	0.291
4	Dieter Fensel	0.293
5	Deborah McGuinness	0.293
6	Jeff Heflin	0.296
7	Frank van Harmelen	0.298
8	Enrico Motta	0.309
9	Guus Schreiber	0.312
10	Raphael Volz	0.317

Figure 4. Top ranking nodes according to aggregate network constraint (structural holes).

There are a number of observations to be made. First, the measures introduced above are distinct from the most common measure of popularity on the web, namely the page count. The sociological measures of closeness and degree centrality and Burt's Structural Holes are only weakly related to the page count, with a correlation of 0.66, 0.45 and 0.51 among the rankings of nodes in the core, respectively. The measure of betweenness is practically unrelated to pagecount, with a correlation of 0.26.

Large discrepancies can be found in both directions. Individuals who are popular on the web, but do not show closeness centrality include Harold Boley, Benjamin Grosof and Jim Hendler. (Taking the entire network into account, we can conclude that from the list of top ten people with the most web pages only three are to be found in the core.) On the other hand, there are also individuals who appear on a relatively small number of pages considering their centrality to the Semantic Web research community (examples include Borys Omelayenko, Andreas Hotho and Ying Ding).

Degree, closeness centrality and Burt's measure are more highly inter-correlated than they are correlated with be-

14 Weak ties are preferred over strong ties as strong ties require time and effort to establish.

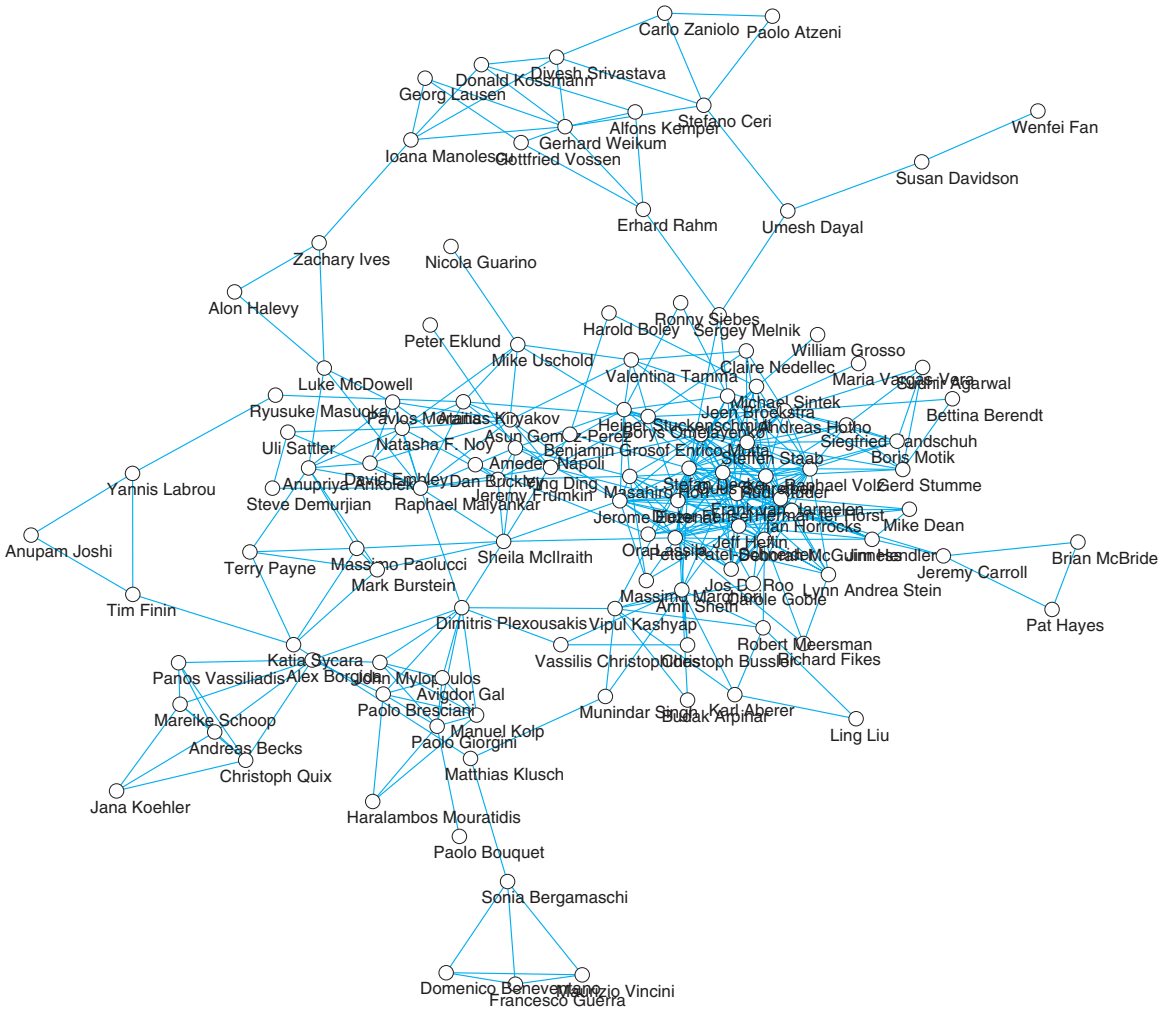

Figure 2. Graphical visualization of the entire network.

tweenness. Betweenness in our network favors those individuals at the outer perimeter of the core who connect the core with large clusters of the periphery. A typical example is the case of Stefan Decker. Decker, having worked in Karlsruhe, have published with various past and present members of the AIFB and DFKI institutes as well as interacting with other key members of the Semantic Web community in highly active projects such On-To-Knowledge and IBROW. This records puts him in a good position to connect the closely knit Karlsruhe-circle with the rest of the network. (See Figure 5.) Burt's Structural Holes strikes a balance among betweenness (which favors the actors on the outer rim of the core) and the measures of degree and closeness centrality, which clearly attribute more power to people in the very center of the core.

Second, there is cognitive evidence that the measures of degree and closeness centrality are successful in distinguishing the important members of the community. 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 and Jeff Heflin, Ian Horrocks and Guus Schreiber have been chairs or the authors of key documents produced by the Web Ontology Working Group of the W3C, the most influential standards organization in the Semantic Web area.

	Degree Centrality		Closeness Centrality		Betweenness Centrality		PageCount (core)		PageCount (complete)	
Rank	Name	Value	Name	Value	Name	Value	Name	Value	Name	Value
1	Andreas Hotho	10	Dieter Fensel	0.480	Stefan Decker	0.251	Jim Hendler	12400	Brian McBride	65400
2	Frank van Harmelen	10	Rudi Studer	0.462	Dieter Fensel	0.224	Ian Horrocks	11800	Pat Hayes	42600
3	Rudi Studer	10	Stefan Decker	0.462	Ying Ding	0.178	Dieter Fensel	10500	Jeremy Carroll	27500
4	Stefan Decker	10	Frank van Harmelen	0.456	Enrico Motta	0.157	Frank van Harmelen	9290	Dan Brickley	26200
5	Dieter Fensel	9	Enrico Motta	0.444	Andreas Hotho	0.137	Stefan Decker	7130	Mike Dean	21900
6	Raphael Volz	9	Guus Schreiber	0.424	Raphael Volz	0.109	Rudi Studer	6920	Jim Hendler	12400
7	Deborah McGuinness	8	Raphael Volz	0.424	Michael Sintek	0.108	Steffen Staab	6870	Stefano Ceri	12100
8	Guus Schreiber	8	Steffen Staab	0.424	Rudi Studer	0.105	Sergey Melnik	6180	Ian Horrocks	11800
9	Ian Horrocks	8	Andreas Hotho	0.404	Jerome Euzenat	0.100	Guus Schreiber	5860	Ora Lassila	10900
10	Jeff Heflin	8	Ian Horrocks	0.404	Frank van Harmelen	0.076	Jeff Heflin	4610	Dieter Fensel	10500

Figure 3. Top ranking nodes according to degree-, closeness- and betweenness centrality and page count.

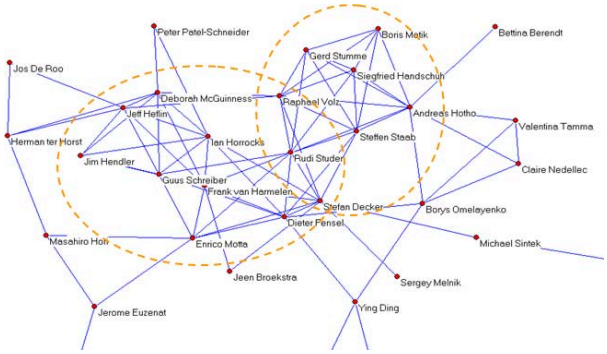

Figure 5. Stefan Decker (along with Rudi Studer and Raphael Volz) plays the role of a local bridge between the Karlsruhe group and other parts of the core.

3.3. Subgroup analysis

As we know from the walks of everyday life, a large community often breaks up to a set of closely knit groups of individuals, woven together more loosely by the occasional interaction across groups. Based on this theory, SNA offers a number of clustering algorithms for identifying communities based on network data. Alternatively, the subgroups may be identified by the researcher using additional attribute data on the subjects. We have tested both approaches in our work.

First, we have used an interactive clustering software provided as a sample with the JUNG Java toolkit for SNA. This software allows the user to cluster a network using an edge-betweenness clusterer and visualize the results. The user can ”play” with the algorithm by adjusting the threshold on a slider-bar. Stronger or weaker clusters appear based on this setting.

Again, the results have clear face validity. As an example, the previously mentioned group of researchers from the AIFB Institute of the University of Karlsruhe quickly emerge as a single cluster of the network. Interesting to note that this affiliation based clustering is less apparent when it comes to groups from other universities, which may be a result of a higher propensity of this group for co-authoring with their colleagues within the AIFB Institute. Clusters based on topicality can also be observed. Members of the Web Services community, for example, can be clearly delineated in the network.

This latter finding gave us the idea of extending our experiment with mining research-topic interests as described in Section 2. As input, we have used a list of 24 key terms which characterize some of the common research directions within the Semantic Web community. We used the Payek tool to analyze the two-mode network that combined the original network with the research-interests of the individuals.

Once again, the results provided clear insight into the community. We translated the two mode network into a regular network where the nodes represent the research topics and the values of the lines give the number of researchers in common to the two topics. (Again, we filtered out the lines below a certain threshold.) The resulting diagram is familiar to those who know the work of this research community (see Figure 6). For example, it is well known that RDQL is the query language of the Jena framework, while the Sesame server implements both RQL and RDQL. It is also the case that OWL is an ontology language, that DAML-S is a Web Services description language with close ties to the agent community and that OWL-S is a successor to DAML-S. Note also that the layout algorithm has a difficulty in placing the concept of Web Services; in reality Web Services is an interdisciplinary topic that recombines many

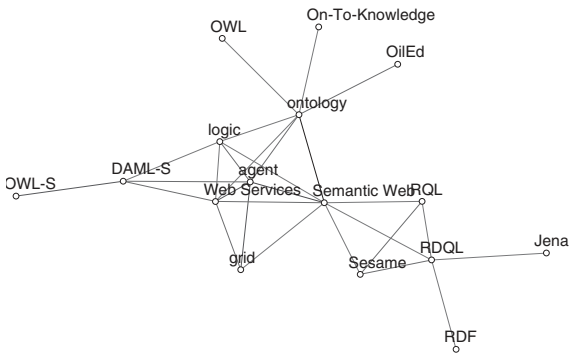

Figure 6. Topic associations with at least 10 people in common.

of the elements of Semantic Web research.

4. Related and future work

The application of Social Network Analysis to web data is one of the most promising methods of social intelligence on the web. In this paper we have presented two advances towards the collection and analysis of information about online social networks. First, we have shown a system for collecting social network data, which combines traditional web mining approaches with crawling the Semantic Web. Second, we have shown how some of the analytical methods of SNA can be applied to the analysis of an online community, namely the core network of Semantic Web researchers.

The scope of related work is far and wide, as our work draws on an interdisciplinary approach combining Web Mining and Social Network Analysis. The closest precursor to our work is the Referral Web project, where data about the AI community was collected in a similar fashion by crawling the web and exploiting a search engine [9]. (Although the data was used for a different purpose, namely interactive exploration of the network and the automation of referrals.) Scientific communities have also been studied by analyzing the linking structure of the Web [1], or relying on publication or project databases [3, 7]. However, none of these earlier work could leverage the Semantic Web, which promises better precision and easier integration of all data sources.

Nevertheless, even by relying on web data we obtained results that have immediate face validity to the researcher who is informed about the workings of the community in question. In the future, we hope to validate this experiment by comparing the results to the outcomes of a survey that uses the traditional questionnaire methods of network analysis. (The validation of an online experiment with real world data is a unique attempt compared to the related work.) We also hope to build more on Semantic Web data in future experiments.

At the same time, we are also aiming to assess how network positions and structures relate to individual and group performance. This kind of analysis is very common in the domain of network research in entrepreneurship [8]. However, traditional studies of entrepreneurship research lack the potential for repeated longitudinal studies due to the effort of manual data collection. Since our system is fully automated at the moment, it gives a so far unexplored opportunity in observing community dynamics over time. We are planning to carry out the data collection throughout a one year period and report in future publications on changes and trends in the structure of the Semantic Web research community.

Although we are only beginning to appreciate the new opportunities that the Semantic Web brings to social intelligence, we hope our work will also inspire others in both the Web Intelligence and Semantic Web communities to explore further the relationships of social structures and the Semantic Web.

References

[1] The EICSTES project (EU-IST-1999-20350): European Indicators, Cyberspace and the Science-Technology-Economy System. http://www.eicstes.org.

[2] Albert-László Barabási. *Linked: The New Science of Networks*. Perseus Publishing, 2002.

[3] A. 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.

[4] D. Brickley and L. Miller. FOAF Vocabulary Specification. Technical report, RDFWeb FOAF Project, 2003.

[5] R. Burt. Structural Holes and Good Ideas (in press). *American Journal of Sociology*, 110(2), 2004.

[6] R. S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1995.

[7] M. Grobelnik and D. Mladenic. Approaching Analysis of EU IST Projects Database. In *Proceedings of the International Conference on Information and Intelligent Systems (IIS-2002)*, 2002.

[8] H. Hoang and B. Antoncic. Network-based research in entrepreneurship: A critical review. *Journal of Business Venturing*, 18:165–187, 2003.

[9] H. Kautz, B. Selman, and M. Shah. The Hidden Web. *AI Magazine*, 18(2):27–36, 1997.

[10] M. Newman. Models of the Small World: A Review. *Journal of Statistical Physics*, 101:819–841, November 2000.

[11] G. Salton. *Automatic text processing*. Addison-Wesley, Reading, MA, 1989.

[12] J. P. Scott. *Social Network Analysis: A Handbook*. Sage Publications, 2nd edition, 2000.

[13] S. Wasserman, K. Faust, D. Iacobucci, and M. Granovetter. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.