Semantic Network Analysis of Ontologies

Bettina Hoser,[1] Andreas Hotho,[2] Robert Jäschke,[2,3] Christoph Schmitz,[2] Gerd Stumme[2,3]

[1] Chair of Informationservices and Electronic Markets, School of Economics and Business Engineering, Universität Karlsruhe (TH), Zirkel 2, D–76128 Karlsruhe, Germany
[2] Knowledge & Data Engineering Group, Department of Mathematics and Computer Science, University of Kassel, Wilhelmshöher Allee 73, D–34121 Kassel, Germany
[3] Research Center L3S, Expo Plaza 1, D–30539 Hannover, Germany

Abstract. A key argument for modeling knowledge in ontologies is the easy re-use and re-engineering of the knowledge. However, beside consistency checking, current ontology engineering tools provide only basic functionalities for analyzing ontologies. Since ontologies can be considered as (labeled, directed) graphs, graph analysis techniques are a suitable answer for this need. Graph analysis has been performed by sociologists for over 60 years, and resulted in the vivid research area of Social Network Analysis (SNA). While social network structures in general currently receive high attention in the Semantic Web community, there are only very few SNA applications up to now, and virtually none for analyzing the structure of ontologies.
We illustrate in this paper the benefits of applying SNA to ontologies and the Semantic Web, and discuss which research topics arise on the edge between the two areas. In particular, we discuss how different notions of centrality describe the core content and structure of an ontology. From the rather simple notion of degree centrality over betweenness centrality to the more complex eigenvector centrality based on Hermitian matrices, we illustrate the insights these measures provide on two ontologies, which are different in purpose, scope, and size.

1 Introduction

A key argument for modeling knowledge in ontologies is the easy re-use and re-engineering of the knowledge. However, beside consistency checking, current ontology engineering tools provide only basic functionalities for analyzing ontologies. Since ontologies can be considered as (labeled, directed) graphs, graph analysis techniques are a promising tool. Sociologists have performed graph analysis since for over 60 years. In the late 1970ies, Social Network Analysis (SNA) emerged as a research area out of this work. Its aim is to analyze the structures of social communities. Typical applications include the analysis of relationships like friendship, communication patterns (e. g., phone call graphs), and the distribution of attendants over several events. While social structures are currently a steeply rising topic within the Semantic Web community (e. g., friend-of-a-friend networks, [4] social tagging systems like del.icio.us.org or www.bibsonomy.org, or semantics-based P2P networks [21]), Social Network *Analysis* has only been applied marginally up to now on ontologies and the Semantic Web.

In this paper, we will discuss the use of SNA for analyzing ontologies and the Semantic Web. While the SNA community has already discovered the internet and the Web as fruitful application domains for their techniques a while ago (e. g., analysing the link structure of the internet [16], and email traffic [17, 22, 25]), SNA applications for the Semantic Web are

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

only emerging slowly. We advocate here a systematic development of *Semantic Network Analyis (SemNA)*, as the adoption of SNA to ontologies and the Semantic Web. In this paper, we show that the application of both basic and advanced SNA techniques to ontologies provide a powerful tool for analyzing the structure of the ontology. We adapt SNA tools to ontology analysis, and discuss the findings. In particular, we discuss how different notions of centrality describe the core content and structure of an ontology. From the rather simple notion of degree centrality over betweenness centrality to the more complex eigenvector centrality based on Hermitian matrices, we illustrate the insights these measures provide on two ontologies, which are different in purpose, scope, and size. The results may be used for selecting the right ontology for a specific application, as well as for re-engineering ontologies.

SemNA is a sub-area of Semantic Web Mining [5]. that addresses the mining of the Semantic Web. To this end, we consider ontologies as (both vertex- and edge-)labeled, directed graphs. As we will discuss below, the existence of different types of nodes and edges (which are reflected in the labels) is a problem for standard SNA approaches. We will discuss solutions for this problem. In this paper, we present two selected applications, and discuss the use of different SNA techniques for analyzing ontologies. The examples will illustrate the deep insights we were able to gain from the two ontologies.

Testcases: SWRC and SUMO ontologies. The SWRC ontology[5] provides a vocabulary about publications, authors, academic staff and the like. It consists of 54 concepts and 70 relations. Figure 1 shows a graphical representation of the ontology. Rectangles represent concepts, relations are shown as rounded boxes.

We selected the SWRC ontology as our first example, as it is a handy size, and as we know its structure rather well, since some of the authors have contributed to its construction. We are thus able to validate the resulting SNA findings (which were computed independently by the non-ontology author) with our insight in the history of the SWRC ontology. The promising results (which were also surprising for the authors) motivated us to consider a larger ontology, the SUMO ontology, where we only knew about its general purpose, but no details about its structure nor its content.

The aim of the Suggested Upper Merged Ontology (SUMO)[6] is to express the most basic and universal concepts for creating a framework for merging ontologies of different domains. With its 630 concepts and 236 relations, SUMO is significantly larger than the SWRC ontology. This information is about all we knew about SUMO when performing our analysis. We are thus in exactly the situation of an ontology engineer who wants to gain deeper insights to a previously unknown ontology.

Organization of the paper. This paper is organized as follows. In the next section, we will provide a brief overview over the history and main lines of research in Social Network Analysis. In Section 3, we will apply a representative selection of SNA techniques to a representative set of ontologies with different structures. In particular, we will analyse the most central parts of the ontology, and will study the eigenvector system assigned to the ontology. Section 4 addresses further applications of SNA for the Semantic Web. In the conclusion, we summarize our experiences, and will discuss the research issues that arise when applying SNA to ontologies and the Semantic Web.

[5] http://ontobroker.semanticweb.org/ontologies/swrc-onto-2001-12-11.oxml
[6] http://www.ontologyportal.org/

2 Social Network Analysis

Already as early as the 1930's Moreno [19] started to describe social relationships within groups using so called *sociograms*. A sociogram is a graph where the members of an observed population are represented as nodes and the relationships among members as edges. The step from modelling relationships between entities of a graph to a structural analysis of these graphs started by using the results from graph theory as early as the 1960's. Pioneers in this field are Harary, Norman and Cartwright [8]. To use the tools of graph theory to analyze and thus describe structures of social networks and to interpret these results in the context of anthropological and sociological contexts was the major achievement of these researchers. The notion of *Social Network Analysis (SNA)* was used to subsume all tools for methodological as well as functional analysis of such group structures.

The two aspects of SNA, the functional aspect and the structural aspect, each highlight a different perspective of research. The functional view focuses on how the function of a network is determined by the structure of a given network. Thus the question of flow between nodes is very prominent. The structural view on the other hand is more interested in the question of structure per se and what statements about a given network can be made based on the analysis of structure alone. Both aspects can be viewed separately, but for some objects of interest, such as organizations, a combined approach may be more appropriate. Since the use of SNA tools in the semantic web environment is just starting out, we will focus in this paper on the structuralist view on SNA, in particular on different notions of centrality. The concept of centrality has many different branches. Just to name a few: in/out degree centrality, betweenness centrality, information centrality, eigenvector centrality. For a good overview see [10].

Wasserman and Faust [26, p.205-219] describe to a great extent the history of rank prestige index, which is an eigenvector centrality based concept. This index is based on the idea, that the rank of a group member depends on the rank of the members he or she is connected to. Stated in mathematical terms this yields the eigenvalue equation (for an eigenvalue equal to 1). The components of the principal eigenvector are the rank prestige indices of each group member. This concept is implemented in the hub-and-authority algorithms of Kleinberg [15] and also in the PageRank algorithm proposed by Page and Brin [7].

There have been different approaches to the analysis of unbalanced graphs. All concepts work very well on undirected and unweighted graphs. But if none of these restrictions apply for a given graph, diffuculties arise. Freeman [11] proposed to use the possibility to split any asymmetric square matrix into its symmetric and skew-symmetric part, perform a singular value decomposition of the skew-symmetric matrix, and showed, that the result could be interpreted as a ranking of dominance. Tyler et al. [25] could identify subgroups in unbalanced email networks by analyzing betweenness centrality in the form of inter-community edges with a large betweenness value. These edges are then removed until the graph decomposes into separate communities, thus re-organizing the graph structure.

Barnett and Rice [3] showed that the transformation of asymmetrical data into matrices that avoid negative eigenvalues may result in the loss of information. This is one of the reasons why we will transform the adjacency matrix into a Hermitian matrix in Subsection 3.3.

Beside considering the direction of links as discussed, the notion of a graph can be refined in several ways. One-mode graphs consider just one type of nodes (e. g., participants of an email network), while two-mode graphs distinguish between two types of nodes (but

still have only one type of edges), forming thus a bipartite graph (e. g., persons and events they are visiting). More general, n-mode graphs distinguish n types of nodes. The edges may also be typed. Extending the definition of [26], we call a *n-mode multigraph with k edge types* a graph where the nodes may be labeled with n different types and the edges with k different labels. This reflects exactly the structure of an (RDFS-based) ontology. Since the interpretation of such complex graphs is more difficult, one often tries to preprocess the data in order to obtain a 1- or 2-mode graph with only one relation, i. e., with one type of edges. Of course the chosen preprocessing transformation has to be taken into account when interpreting the results.

To analyze networks more easily, several software tools have been developed. These packages include, but are not limited to UCINet,[7] Pajek,[8] and Visone.[9] There are also packages for R[10] and also some implementations in Java.[11] For a good overview on SNA and its history refer to Wasserman and Faust [26] and Freeman [12].

3 Network Analysis of Ontologies

Ontologies can be considered as n-mode multi-graphs with k edge types. As argued above, n-mode multi-graphs with k edge types are hard to analyze when n is larger than 2 or 3, and k is larger than 1. Therefore we follow the usual approach of projecting them first to a 1- or 2-mode 1-plex network. In the sequel of this section, we will first illustrate the benefit of some basic SNA approaches, before performing a more sophisticated analysis, based on the analysis of the eigenvectors of the adjacency matrix. To show the diversity of results that can be expected from such an analysis, we will apply the basic techniques to two different ontologies: the SWRC and the SUMO ontology, which differ in purpose, scope, and size.

3.1 Preprocessing the Ontologies

As SNA techniques work on graphs, we first have to transform the ontology into a suitable graph. As in all knowledge discovery (KDD) applications (and probably even more as in the average KDD scenario), the interpretation of the final results is highly sensitive to the decisions made during this preprocessing step.

A standard approach (which we will follow as well) of turning n-mode networks with k edge types into a (directed or undirected) graph is to collect all types of nodes into just one set of nodes, and to ignore the edge types.[12] We will keep the typing information, though, and refer to it during the analysis. As a first step, we set up a directed graph for the input ontology in the following way:

Technical artifacts were pruned from the ontology. In the KAON ontology API, [13] which was used in our experiments, these comprise the artificial root concept which is present in

[7] http://www.analytictech.com/ucinet.htm
[8] http://vlado.fmf.uni-lj.si/pub/networks/pajek/
[9] http://www.visone.info/
[10] http://www.stat.ucl.ac.be/ISdidactique/Rhelp/library/sna/html/00Index.html
[11] http://jung.sourceforge.net/index.html
[12] A more frequent way for handling different edge types is to perform a sequence of analyses, one for each edge type. For ontologies, however, this approach is not suitable, as most edge types (beside 'is_a' and eventually 'part_of') appear only once.
[13] http://kaon.semanticweb.org/

all ontologies, and entities for lexical information such as labels and word stems. Each concept and each property became a node in the graph. Between two concepts C_1 and C_2, a directed edge (C_1, C_2) was added if C_1 is a direct subconcept of C_2. Between concept and property nodes, edges are added as follows: an edge is added from each domain concept node to its property nodes, an edge is added from each property node to its range concept node, unless the property was scalar-valued or untyped, and an edge is added from each subproperty to its superproperty.

The adjacency matrix A of this graph has one row and one column for each node. If there is an edge from the ith to the jth node, then $a_{ij} := 1$, else $a_{ij} := 0$. This matrix is the subject of our subsequent analysis. For the SWRC ontology, A has thus 54+70 rows and 54+70 columns, with entries 0 and 1. The matrix for the SUMO ontology is structured in the same way, with $630 + 236 = 866$ rows and columns in total.

3.2 Basic Methods of Network Analysis

The intuitive approach to analyze a network, represented as a graph $G := (V, E)$ with nodes (or vertices) $v \in V$ and edges $e \in E$, is to start with the number of connections each node has. A node that has many connections is presumed to be important, while a node without connections is presumed to be irrelevant. This concept is called *degree centrality*. In the adjacency matrix A the degree centrality c_k of a vertex in an undirected graph can be calculated as the row or column sum $c_k = \sum_l^n a_{kl}$ of A. If the connection between two nodes has no directional preference this is just called *degree*. If the relationship has an inherent direction, like in 'person A called person B' then the degree is categorized into *in-* (column sums) and *outdegree* (row sums) depending on whether the connection ends at a node or starts at a given node.

The *betweenness centrality* is the (normalized) number of shortest paths between any two nodes that pass through the given node. The betweenness centrality provides often a high degree of information, as it describes the location of a node in the graph in a global sense, while in- and outdegree consider the direct neighbor nodes only.

Based on the degree centrality we can define the *density* d of a network. Let the network describe a non-directional relationship between nodes, then the density is defined as the number of existing connections divided by the number $N := \frac{|V|(|V|-1)}{2}$ of all possible edges as $d = \frac{\sum_{kl} a_{kl}}{N}$. Thus a completely connected network has a density of 1. In the directed case one has to keep in mind that at most two connections are possible between two nodes. Thus the density d_d becomes $d_d = \frac{\sum_{kl} a_{kl}}{|V|(|V|-1)}$. This concept is not useful anymore when multiple connections are allowed or when the connections become valued or weighted, because no total number of possible connections can be given in that case.

Another measure of how well a graph is connected is its *diameter*. For all pairs A, B of nodes, we calculate the shortest path from A to B, and take then the maximum over their lengths. The well-known *small-world phenomenon* states that social networks have a small diameter. Diameter and density are used for comparing networks.

Global comparison of SWRC and SUMO. To analyze the given ontologies, we calculated for each of them the diameter and the density of the network. The results are shown in Table 1. These indices were generated using Pajek.

Compared to typical social networks, the density of the SWRC ontology (0.015) is very sparse. SUMO has an even sparser density with 0.0024. The fact that the difference between

	# concepts	# relations	diameter	density
SWRC	54	70	16	0.015
SUMO	630	236	27	0.0024

Table 1. Size, diameter, and density of SWRC and SUMO

both ontologies is approx. one magnitude, which is in the same ratio as their difference in size, indicates that the concepts in both ontologies have a similar number of properties attached in average. It might be interesting to analyze more ontologies to check whether this is some kind of constant stemming from ontology engineering principles. We assume that ontologies are scale-free networks because of their construction.

For studying both ontologies in more details, we computed as next step for all their nodes indegree, outdegree and betweenness centrality.

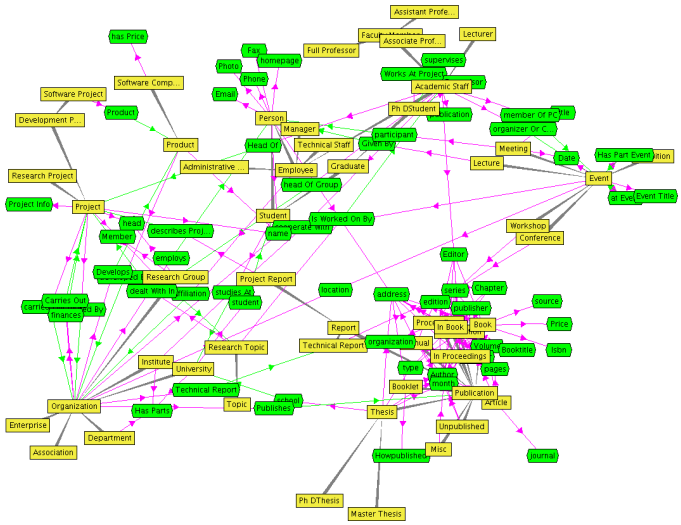

Fig. 1. The SWRC Ontology

The SWRC ontology in detail. Table 2 shows the indegrees, outdegrees and betweenness centralities of the nodes in the graph extracted from the SWRC ontology. While the degrees could still be read from Fig. 1, the betweenness centrality has to be listed.

Considering the degrees only, one observes that the BibTeX part of the ontology was modeled with the highest level of detail: BibTeX-related concepts such as ‘Book’ and ‘InCollections’ have high outdegrees (i. e. a large number of properties) but no indegree, while the related properties such as ‘author’, ‘month’, and ‘address’ have large indegree.

Properties which apply to all kinds of publications, such as ‘title’ and ‘year’, have a low degree, as they are attached to ‘Publication’ only and are inherited by its subclasses. This is a result of the way we set up the adjacency matrix. An alternative way of setting up the matrix is to model explicitly also the inherited attributes. This is an example for the fact that the modeling step has to be taken into account for the interpretation of the SNA results.

#	Label	d_o	d_i	b_C
1	Academic Staff	10	4	0.102
2	Administrative Staff	1	0	0.
3	Article	7	0	0.
4	Assistant Professor	1	0	0.
5	Associate Professor	1	0	0.
6	Association	1	0	0.
7	Book	13	0	0.
8	Booklet	5	0	0.
9	Conference	2	0	0.
10	Department	2	0	0.
11	Development Project	1	1	0.004
12	Employee	2	4	0.025
13	Enterprise	1	0	0.
14	Event	6	9	0.019
15	Exhibition	1	0	0.
16	Faculty Member	1	3	0.013
17	Full Professor	1	0	0.
18	Graduate	1	1	0.016
19	In Book	13	0	0.
20	In Collection	14	0	0.
21	In Proceedings	12	0	0.
22	Institute	3	0	0.
23	Lecture	2	0	0.
24	Lecturer	1	0	0.
25	Manager	1	0	0.
26	Manual	6	0	0.
27	Master Thesis	1	0	0.
28	Meeting	4	1	0.001
29	Misc	3	0	0.
30	Organization	8	10	0.134
31	Person	7	5	0.024
32	PhDStudent	4	1	0.024
33	Ph DThesis	1	0	0.
34	Proceedings	9	0	0.
35	Product	2	3	0.017
36	Project	7	7	0.12
37	Project Meeting	1	0	0.
38	Project Report	2	0	0.
39	Publication	5	14	0.022
40	Report	2	2	0.005
41	Research Group	3	1	0.01
42	Research Project	1	1	0.004
43	Research Topic	3	1	0.079
44	SoftwareComponent	2	0	0.
45	Software Project	2	0	0.
46	Student	2	3	0.027
47	Technical Report	3	1	0.014
48	TechnicalStaff	1	0	0.
49	Thesis	6	2	0.01
50	Topic	1	1	0.
51	Undergraduate	1	0	0.
52	University	3	2	0.041
53	Unpublished	3	0	0.
54	Workshop	2	0	0.
55	Abstract	0	1	0.
56	address	0	9	0.
57	Affiliation	1	1	0.019
58	AtEvent	1	1	0.
59	author	0	10	0.
60	booktitle	0	2	0.
61	carried Out By	1	1	0.009
62	carriesOut	1	1	0.033
63	Chapter	0	2	0.
64	cooperate With	0	2	0.
65	Date	0	2	0.
66	Dealt With In	1	1	0.004
67	Describes Project	1	1	0.004
68	Developed By	1	1	0.017
69	develops	1	1	0.006
70	edition	0	4	0.
71	editor	0	6	0.
72	Email	0	1	0.
73	employs	1	1	0.013
74	Event Title	0	1	0.
75	fax	0	1	0.
76	financedBy	1	1	0.009
77	Finances	1	1	0.033
78	Given By	1	1	0.
79	Has Part Event	1	1	0.
80	Has Parts	0	3	0.
81	hasPrice	0	1	0.
82	head	0	2	0.
83	Head Of	1	1	0.011
84	head Of Group	1	1	0.008
85	homepage	0	1	0.
86	howpublished	0	2	0.
87	institution	0	0	0.
88	Is About	1	1	0.078
89	IsWorkedOnBy	1	1	0.069
90	Isbn	0	1	0.
91	Journal	0	1	0.
92	Keywords	0	1	0.
93	Location	0	2	0.
94	member	0	2	0.
95	member Of PC	1	1	0.01
96	month	0	11	0.
97	name	0	6	0.
98	Note	0	1	0.
99	number	0	6	0.
100	organization	0	4	0.
101	organizer Or Chair Of	1	1	0.01
102	Pages	0	4	0.
103	participant	1	1	0.001
104	phone	0	1	0.
105	Photo	0	1	0.
106	Price	0	1	0.
107	product	1	1	0.001
108	projectInfo	1	1	0.009
109	publication	0	2	0.
110	Publisher	0	5	0.
111	publishes	1	1	0.01
112	School	1	1	0.012
113	series	0	8	0.
114	source	0	1	0.
115	has student	1	1	0.002
116	Studies At	1	1	0.024
117	Supervises	1	1	0.023
118	supervisor	1	1	0.006
119	TechnicalReport	1	1	0.017
120	Title	0	2	0.
121	Type	0	3	0.
122	Volume	0	6	0.
123	Works AtProject	0	2	0.
124	Year	0	1	0.
	Mean (Degree)	1.82	1.82	–
	Std (Degree)	2.84	2.55	–

Table 2. Degree and betweenness centrality of concepts (# 1–54) and relations (# 55–124)

The betweenness centrality gives us a more global description of the roles of nodes in the graph. For SWRC, it returns first of all 'Organization' and 'Project', followed in short distance by Academic Staff' and 'Research Topic'. These are thus the concepts that play a 'bridging role' in SWRC; they are used for describing (chains of) other objects (these are the incoming edges), and they are described by (chains of) other objects (the outgoing edges). From a database perspective, these are typical candidates for joins in a query.

The SUMO ontology in detail. We also computed the list of in, out and between degrees of the concepts and relations of the SUMO ontology. Due to space restrictions, we omit this list. The means of in- and outdegree (which are obviously equal, as each outgoing edge has to go in somewhere) are at 2.07. The standard deviation is 1.67 for the outdegrees, and 5.8 for the indegrees. The large difference of the standard deviations indicates a heterogeneity in the way of modeling.

When looking at the concepts and relations with out- and indegrees differing largely from the mean, this heterogeneity can be explained. The highest indegree has the concept 'BinaryPredicate' ($d_i = 102$), and the highest outdegree has the concept 'Process' ($d_o = 20$). The former shows that this technical notions is important for the designers of the ontology. However, this concept is conceptually not part of the domain of interest of the ontology, but rather a meta-construct. If the KR language permitted different arity relations, this would be modeled with language constructs and not by reification. The latter, on the other hand, indicates that 'Process', which is indeed a concept of the domain of interest, is modelled in a high level of detail by providing many properties that a process can have. As

	Outdegree	Indegree
Process	20	10
Object	15	21
RealNumber	13	15

	Outdegree	Indegree
BinaryObject	3	102
AsymmetricRelation	2	71
UnaryFunction	3	54

Table 3. Highest out- and indegrees of SUMO concepts.

in the SWRC ontology, the betweenness centrality emphasizes more on the conceptual part of the ontology: the top node according to this measure is 'Object', followed by 'Formula', 'Entity', 'Physical', 'List', 'Process'. These are the central nodes of the SUMO ontology.

3.3 Eigensystem Analysis

Compared to the centrality measures described so far, the eigensystem of the adjacency matrix provides an overall view of the network, while still allowing a very detailed structure analysis of its parts.

Eigenvector centrality measurements have become a standard procedure in the analysis of group structures. Mostly symmetric (dichotomized) data has been used. Bonacich and Lloyd [6] present an introduction of the use of eigenvector-like measurements of centrality for asymmetric data. The analysis of directed, weighted, asymmetric relationships within a social network poses some difficulties. In this paper we will use a method based on the status (rank prestige) index method [26, p.205-219], that was adapted by the first author to complex adjacency matrices. We sketch the principal approach here (the technical details are presented in [14] and [2]) and adapt it to the analysis of ontologies.

In the following, we consider an ontology as a network which can be modeled as a directed, weighted graph $G = (V, E)$ with V denoting the set of nodes or members and E denoting the set of edges, links or communications between different members. Self references (loops) are excluded.

We use the following construction rules for a complex adjacency matrix H of the initial graph G: First, we construct a square complex adjacency matrix C with n nodes from the possibly weighted real valued adjacency matrix A of graph G by $C = A + iA^t$ with $a_{kl} = m + ip$ where m is the number of outbound edges (or equivalently the weight of the outbound edge) from node k to node l, p is the number of inbound edges (or equivalently the weight of the inbound edge) from node l to node k, and i is representing the imaginary unit ($i^2 = -1$). As can be seen, $c_{kl} = i\overline{c_{lk}}$ holds. Then we rotate C by multiplying it with $e^{-i\frac{\pi}{4}}$ in order to obtain a Hermitian matrix H, i. e., $H := C \cdot e^{-i\frac{\pi}{4}}$. For the proof see [2].

The fact that the resulting matrix is Hermitian has the advantage that it has full rank and thus a complete orthogonal eigenbasis can be found. The consequence is that H can be represented by a Fourier sum as the sum of all orthogonal projectors $P_k = \mathbf{x_k}\mathbf{x_k^*}$, weighted by the corresponding eigenvalue λ_k: $H = \sum_{k=1}^{n} \lambda_k P_k$. Since all eigenvalues are real, they can be sorted by absolute value. In addition the eigenvalue can be used to calculate the covered data variance. These characteristics can be used to analyze a network structure at different levels of relevance as will be shown later in this paper.

Under this similarity transformation the coordinate independent characteristics of the original directional patterns are kept, no information is lost. For instance, more outbound than inbound links lead to a negative sign of the imaginary part of h_{kl}, while more inbound than outbound links lead to a positive sign of the imaginary part of h_{kl}. Now one can analyze the eigensystem of the matrix H in order to gain insights into the structure of the underlying ontology.

Eigensystem analysis of the SWRC ontology. We start by using the adjacency matrix A for the SWRC ontology from subsection 3.2, and construct the matrix H as described above. This matrix is the subject of further examination.

Let us first have a look at the distribution of the eigenvalues of H as shown in Fig. 2. The diagram suggests a symmetry in the spectrum. This indicates that major components of the network are star like in structure. As the concept hierarchy of SWRC is a tree, this hierarchy has a snowflake structure if considered as graph. Hence our observation that stars are predominant indicates that the concept hiararchy has a more important influence on the overall structure of the SWRC ontology than the non-hierarchical relationships.

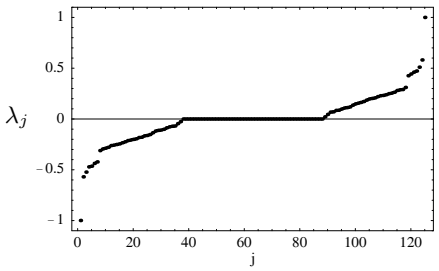

Fig. 2. Eigenspectrum of the ontology sorted by value

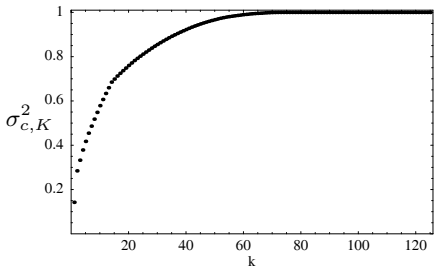

Fig. 3. Cumulative covered variance $\sigma^2_{c,K}$ of the SWRC ontology by eigenvalues λ_k

Fig. 3 displays the cumulative covered variance of the ontology. One can see that the first two eigenvalues cover already 29 % of the variance of the system, that it has a clear distance to the following eigenvalue, and that the first 14 eigenvalues cover approx. 70 % of the overall variance. The remaining eigenvalues contribute marginally only.

In Fig. 4 we now take a more detailed look at the eigenvectors and their components. The lefthand side gives the eigenvalues of each eigenvector, the righthand side gives covered data variance, each eigenvector is represented horizontally with the components numbered 1 through 125 on the bottom, and each eigenvector component is represented as a colored (or gray scaled) field.

The eigenvector components are complex valued, indicating in the phase of the complex number the direction of the connection with respect to the central node,and in the absolute value the relevance of the node in this eigenvector. The color representation lends itself naturally. The absolute value of the component is given by the brightness of the colored field. In gray scales an absolute value of 0 or near 0 is black, while an absolute value close to 1 is bright or has a saturated color. The phase of the complex number is represented by color where a phase of 0 is given as red and counter clockwise $\frac{\pi}{4}$ is yellow, $\frac{\pi}{2}$ is yellow-green, $\frac{3}{4}\pi$ green, $-\pi$ cyan, $-\frac{3}{4}\pi$ blue, $-\frac{\pi}{2}$ blue-magenta, $-\frac{\pi}{4}$ magenta and coming back to red. Thus for example the field with the coordinates $1.0, 39$ is bright red which indicates an eigenvalue with high absolute value and phase 0.

By checking for the largest eigenvector component in each of the eigenvectors (colored red) corresponding to these eigenvalues we can see which concept/relation of the ontology is most central: In the first eigenvector (i. e., the lowest row in Fig. 4, with eigenvalue $+1$), the brightest color is in column 39, which is the concept 'Publication'. The fact that the same column shows in the eigenvector for the negative of the eigenvalue (i. e., in the second row from below, with eigenvalue -1) the same phase (as it is red as well) indicates that the

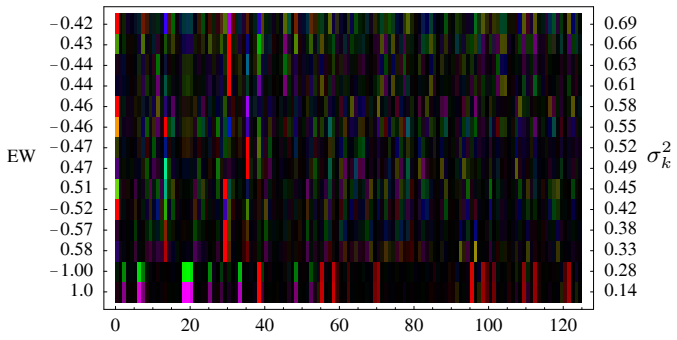

Fig. 4. The 14 strongest eigenvectors of the ontology

concept 'Publication' is the center of a star like structure. The concept 'Organization' (= column 29) follows (at some distance) with the third and fourth eigenvector. This confirms that publications were in the key focus of the developers of SWRC – a finding we were already pointed to when analysing the in- and outdegrees in the previous subsection. In fact, this fits with the history of the development of the SWRC ontology, which started by transforming the BibTeX format into an ontology.

When looking further down the eigenvalues, we observe that of the three concepts 'Academic Staff', 'Employee' and 'Person', 'Academic Staff' already becomes relevant in the fifth eigenvector, while 'Person' becomes relevant as late as the 11th eigenvector. 'Employee' does not feature as a central concept in any eigenvector. This observation raises the question if the concepts 'Employee' and 'Person' are really needed by the applications the SWRC ontology is targeted to, or if they eventually have just been added because 'one is usually doing so' when designing an ontology.

In Fig. 4, we observe also that the concept 'Academic Staff' *interlaces* with 'Organisation', 'Project' and 'Person'. This behavior is visible by observing that while 'Academic Staff' is colored red in the fifth eigenvector (eigenvalue -0.52), it changes color already in the next line and goes back to red again in line 10 and again in line 14. The three other concepts are colored red in the remaining eigenvectors in between. The absolute values of the eigenvalues do not come in strict pairs of equal absolute value but different sign, thus the three star like structures can not be clearly separated into blocks. The pattern of connection of AcademicStaff to the rest of the network is not easily explained. The pattern of AceademicStaff is distrubed by other structures that have approximately the same amount of connections, thus seperating the eigenvalues.

When considering the eigenvectors of the 36th to 44th eigenvalue (which are out of Fig.4 due to space restrictions), we observe that the concepts 'Assistant Professor', 'Associate Professor', and 'Full Professor' (columns 4, 5, and 17) behave identically with respect to 'Faculty Member' (column 16). As these three concepts are also very similar from an ontology engineering point of view, we take this as a hint that, in a re-engineering step, they should be unified to a single concept, with an additional attribute like 'status'.

As a last view we take a look at the partial sums as described earlier. In Fig. 5 we see the partial sum of the Fourier sum of the first two eigenprojectors weighted by their eigenvalues and rotated back ($\sum_{k=1}^{2} \lambda_k P_k$). This figure was generated by using an adapted k-means cluster algorithm based on the eigensystem. To define the initial cluster centers

Fig. 5. Back rotated partial sum of first two eigenprojectors

we use the eigenvector components with the highest absolute value of those eigenvectors that have a negative eigenvalue. We further restrict the selection to all those eigenvectors where the eigenvalues add up to explain data variance to a predefined level of 70%. Thus we do not need to set the number of clusters ex ante. An approximated block matrix is generated when we then sort the eigenvectors and rearrange the eigenvector components accordingly before calculating the eigenprojector. Since the matrices are hermitian, the blocks are symmetric but different in color. The color-coding is the same as in Fig. 4. What is clearly visible is the BibTeX structure as a block in the upper left hand corner. It shows a very strong outbound connection from concepts like 'Book', 'InBook', etc. to Publication', 'address' and 'edition' for example.

If we now take the partial sum of the first 14 eigenprojectors we bring more detail to the picture. In Fig. 6 we see in addition to the BibTeX block five right angles in the matrix plot. These five structures belong to the concepts of 'Organisation', 'Academic Staff', 'Project', 'Event' and 'Person'. As this matrix can be read as a 'partial adjacency matrix', such right angles are the structure one expects for stars in the graph: one central node pointing from/to several nodes around it. Different to the BibTeX block that is visible in the upper left hand corner, these concepts play thus a central role in their surroundings. The color of the horizontal part of the angle indicates the direction: for 'Organization', it is green, hence this concept has many inbound edges – its subconcepts. The red color for 'Acadamic Staff' comes from its many outbound properties. 'Project', 'Event' and 'Person' have both incoming and outcoming edges/properties.

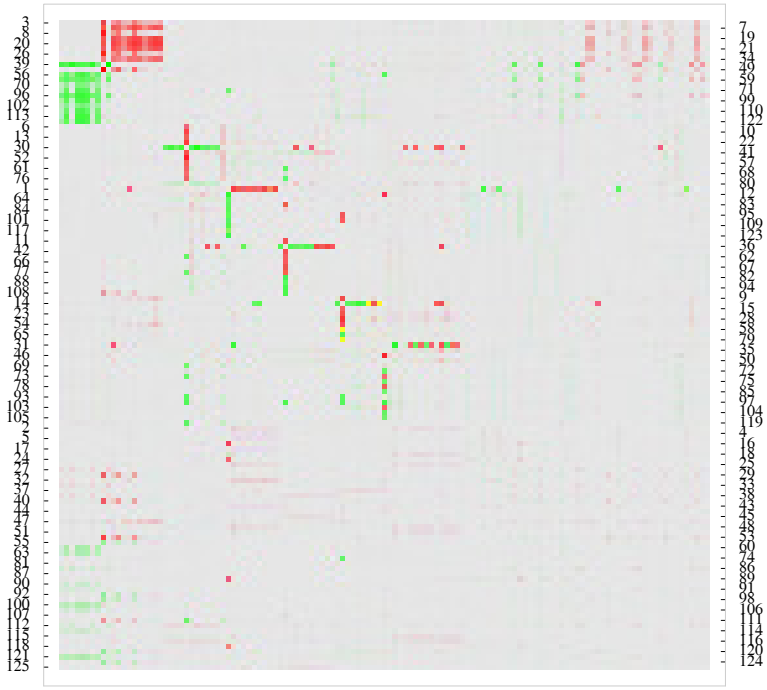

Fig. 6. Back rotated partial sum of first 14 eigenprojectors

Eigensystem analysis of the SUMO ontology. The eigensystem of the SUMO ontology differs significantly from the one of SWRC. Not only because the SUMO ontology is modeled as a graph with more then 800 nodes, but if differs in that this ontology does not have such a very prominent center.

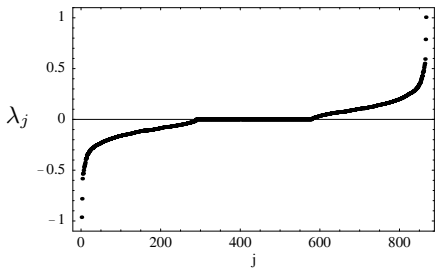

Fig. 7. Eigenspectrum of the SUMO ontology sorted by value

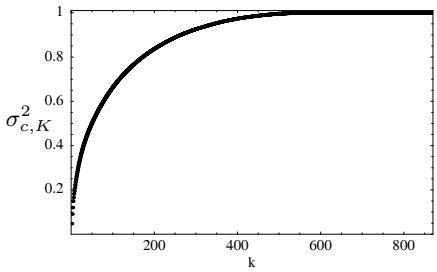

Fig. 8. Cumulative covered variance $\sigma^2_{c,K}$ of the SUMO ontology by eigenvalues λ_k

The spectrum of SUMO (given in Fig. 7) shows – as in the SWRC case – a very strong symmetry, thus suggesting star like structures which come again from the concept hierarchy where several subconcepts all point to their common superconcept. Different to SWRC, the cumulative covered variance (Fig. 8) shows a rather slow incline. While the first two

eigenvalues of the SWRC ontology covered already 29% of the data variance, the first two eigenvalues of SUMO cover only about $10\,\%$. The incline then goes without any obvious steps. This suggests that many concepts need to be taken into account to explain the complete ontology. Otherwise said, the degree of detail in SUMO seems to be more balanced than in SWRC.

Due to space restrictions, we cannot display the equivalents of Figs. 4 to 6 for SUMO here. We only present the major insights of our analysis verbally. The concept 'Binary Predicate' contributes most to the interpretation of the first two eigenvectors. 'Asymmetric Relation' seems to follow the same pattern in connecting to other nodes. Thus it is the second strongest concept in the first two eigenvectors. The fact that these two concepts also have a high absolute value in the following six eigenvectors further indicates that these two concepts also contribute to a high extent to the interpretation of these patterns. This might tells us that, in SUMO, these two concepts play a predominant role.

The third and fourth eigenvectors are most strongly influenced by the concepts 'Unary Function', 'Total Valued Relation' and 'Unit of Measure'. These three concepts have similar incoming connections from many concepts which are all of the form'...Fn'. This can be taken as a hint that these bundles of relations could be unified if there were a suitable construct in the KR formalism.

Concluding this section, we summarize that the out-/indegree analysis (and in particular the different differences of the standard deviations for out- and indegree) showed us that SUMO is more heterogenous in its way of modeling (due to the lack of a construct for higher-arity relations in the KR language) than SWRC, but that it is – according to the eigensystem analyis – more homogenous in the distribution of the coverage of different sub-domains of interest.

4 Other Applications of SNA in the Semantic Web Context

The application of SNA in the Semantic Web context is emerging only right now, but there are interesting first results. In [18], Mika defines a model of semantic-social networks for extracting lightweight ontologies from the upcoming field of folksonomies. Besides calculating such measures like the clustering coefficient, (local) betweenness centrality or the network constraint on the extracted one-mode network, Mika uses co-occurence techniques for clustering the concept network. Stuckenschmidt [23] uses network analysis to partition an ontology into a disjoint and covering set of concepts. After creating a dependency graph of the ontology and computing the strength of the dependencies the line island method [4] is used to determine strongly related concepts. These are then used to form a partition of the ontology graph. The tool Ontocopi described in [1] performs what is called Ontology Network Analysis for initially populating an organizational memory. Several network analysis methods are applied to an already populated ontology to extract important objects. In particular, a PageRank-like [7] algorithm is used to find communities of practice of individuals represented in the ontology.

Another field of interest regarding SemNA are Friend Of A Friend (FOAF) [14] networks which are studied for instance in [20] and [9]. Both articles focus on analysing the structure of the social network yielded by a large collection of FOAF documents.

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

5 Conclusion

In this paper, we have shown that Social Network Analysis provides a promising set of tools for analyzing ontologies and Semantic Web applications, providing deep insights into the structure of ontologies and knowledge bases. In particular, we have seen that the analysis of a given ontology can be done very thoroughly at different levels of granularity. The gained insights may help to design or redesign ontologies in such a way as to find redundancies or holes that should be mended. The analysis is also of use for selecting the right ontology for reuse. Consider the case where, for a specific purpose, a search (for instance with Swoogle[15] returned several ontologies. The eigenvalue analysis provides deep insights into the structure and focus of the ontology and supports the selection of the most suitable result.

As the two research areas Semantic Web and Semantic Network Analysis met only recently, open issues are still abundant, and provide a rich domain of research for the coming years. We will conclude by discussing some of these upcoming issues.

- As seen above, SNA deals well with one- to n-mode networks with one relation. However, ontologies typically consist of more than one or two concepts, and of more than just one kind of relation. A systematic analysis of preprocessing steps which transform an ontology into a one-/two-mode simple structure is thus needed. This has to take into account a deep understanding of how the transformations effect the interpretability of the results.
- One step further in this direction is the very interesting — and far from trivial — research question how to expand existing SNA approaches to n-mode multigraph data sets.
- The analysis of the results of the standard eigenvector analysis needs currently some experience. Future work includes the use of cluster algorithms for rearranging the dimensions of the vector space such that similar dimensions are visualized together.
- (Description) Logics based ontologies describe relations (such as the subsumption hierarchy) implicitly only. It has to be studied in how far these relations have to be computed explicitly before SNA techniques can be applied in a meaningful way.
- The next step after analyzing the ontologies is to turn the outcome into support for search, navigation, browsing, and restructuring ontologies and knowledge bases. Some attempts have been made, as mentioned above. Seen the large field of SNA techniques, though, we expect a lot more techniques and tools to come up within the next years.
- Another direction of research is the comparison with philosophical aspects of ontology engineering. The OntoClean [13] method provides a framework for the evaluation of ontological decisions bsaed on philosophical notions e.g. of Identity or Polysemy. Correlations between the structural and philosophical properties of ontologies will have to be researched.

References

1. Harith Alani, Srinandan Dasmahapatra, Kieron O'Hara, and Nigel Shadbolt. Identifying Communities of Practice through Ontology Network Analysis. *IEEE Intelligent Systems*, 18(2):18–25, March/April 2003.

[15] http://swoogle.umbc.edu/

2. A. Geyer-Schulz B. Hoser. Eigenspectralanalysis of Hermitian Adjacency Matrices for the Analysis of Group Substructures. *Journal of Mathematical Sociology*, 29(4):265–294, 2005.
3. George A. Barnett and Ronald E. Rice. Longitudinal non-euclidean networks: Applying galileo. *Social Networks*, 7:287–322, 1985.
4. Vladimir Batagelj. Analysis of large networks - Islands. Presented at Dagstuhl seminar 03361: Algorithmic Aspects of Large and Complex Networks, August/September 2003.
5. B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In *Proc Int. Semantic Web Conference*, Sardinia, Italy, 2002.
6. P. Bonacich and P. Lloyd. Eigenvector-like measurement of centrality for asymmetric relations. *Social Networks*, 23:191 – 201, 2001.
7. Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.
8. Frank Harary ; Robert Z. Norman ; Dorwin Cartwright. *Structural models : an introduction to the theory of directed graphs*. Wiley, New York, 1965.
9. Li Ding, Lina Zhou, Timothy W. Finin, and Anupam Joshi. How the Semantic Web is Being Used: An Analysis of FOAF Documents. In *HICSS*. IEEE Computer Society, 2005.
10. M.G. Everett and S.P. Borgatti. The centrality of groups and classes. *Journal of Mathematical Sociology*, 23(3):181–201, 1999.
11. Linton C. Freeman. Uncovering organizational hierarchies. *Computational & Mathematical Organization Theory*, 3(1):5 – 18, 1997.
12. Linton C. Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. BookSurge Publishing, 2004.
13. Nicola Guarino and Christopher A. Welty. Evaluating ontological decisions with OntoClean. *Commun. ACM*, 45(2):61–65, 2002.
14. Bettina Hoser. *Analysis of Asymmetric Communication Patterns in Computer Mediated Communication Environments*. PhD thesis, Universität Karlsruhe, 2005.
15. Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Ninth Annual ACM-SIAM Symposium*, pages 668 – 677, Jan 1998.
16. Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *JACM*, 46(5):604–632, sep 1999.
17. Barry Wellman Laura Garton. Social impacts of electronic mail in organizations: A review of research literature. *Communication Yearbook*, 18:434–453, 1995.
18. Peter Mika. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In Yolanda Gil, Enrico Motta, V. Richard Benjamins, and Mark A. Musen, editors, *ISWC 2005*, volume 3729 of *LNCS*, pages 522–536, Berlin Heidelberg, November 2005. Springer-Verlag.
19. J.L. Moreno. *Who shall survive? : a new approach to the problem of Human Interrelations*, volume 58 of *Nervous and mental disease monograph series*. Nervous and Mental Disease Publ., Washington, 1934.
20. John C. Paolillo, Sarah Mercure, and Elijah Wright. The Social Semantics of LiveJournal FOAF: Structure and Change from 2004 to 2005. In Stumme et al. [24].
21. Christoph Schmitz. Self-organization of a small world by topic. In *Proc. 1st International Workshop on Peer-to-Peer Knowledge Management*, Boston, MA, August 2004.
22. Michael F. Schwartz and David C. M. Wood. Discovering Shared Interests Using Graph Analysis. *Communications of the ACM*, 36(8):78 – 89, Aug 1993.
23. Heiner Stuckenschmidt. Network Analysis as a Basis for Ontology Partitioning. In Stumme et al. [24].
24. Gerd Stumme, Bettina Hoser, Christoph Schmitz, and Harith Alani, editors. *Proc. ISWC 2005 Workshop on Semantic Network Analysis (SNA2005)*, Galway, Ireland, November 2005.
25. Joshua R. Tyler, Dennis M. Wilkinson, and Bernardo A. Huberman. Email as spectroscopy: Automated discovery of community structure within organizations. *cond-mat/0303264*, 2003.
26. Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, volume 8 of *Structural Analysis in the Social Sciences*. Cambridge University Press, Cambridge, 1 edition, 1999.